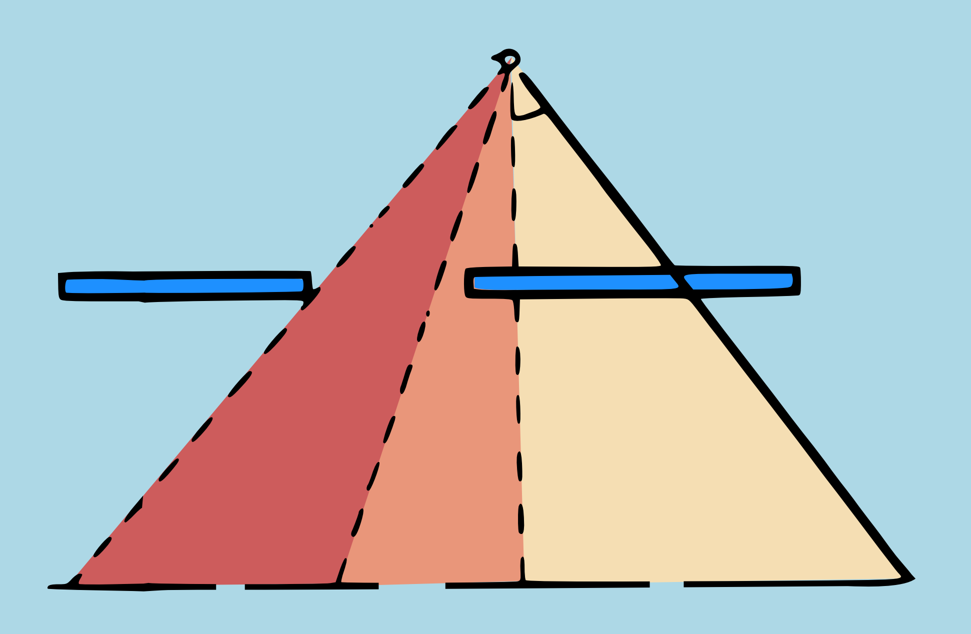
A. Borovkov

STATISTIQUE MATHÉ-MATIQUE



Éditions Mir Moscou

А. А. БОРОВКОВ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ИЗДАТЕЛЬСТВО «НАУКА»
МОСКВА

A. BOROVKOV

STATISTIQUE MATHÉMATIQUE



ÉDITIONS MIR MOSCOU

Traduit du russe
par DJILALI EMBAREK

На французском языке

TABLE DES MATIÈRES

AVANT-PROPOS

Cet ouvrage s'inspire des cours de statistique mathématique professés durant de longues années par l'auteur aux élèves de troisième année de la faculté de mathématiques de l'université de Novossibirsk. L'auteur a modifié le contenu à plusieurs reprises en quête de la version qui soit la plus élégante, la plus accessible et à jour. Plusieurs variantes ont été essayées, à commencer par l'exposition, sous forme de recettes, des principaux types de problèmes (construction des estimateurs et des tests et étude de leurs propriétés) pour finir par un cours de théorie générale des jeux dans lequel la théorie des estimateurs et le test d'hypothèses ont été présentés comme des cas particuliers d'une même approche. Le manque de temps (un semestre) n'a pas permis de regrouper ces deux variantes qui se complètent malgré leurs imperfections évidentes. Dans le premier cas, certains faits concrets ont entravé le développement d'une vision globale de la matière étudiée. La deuxième variante pour sa part comportait peu de résultats concrets simples et beaucoup de notions nouvelles difficiles à l'assimilation. En tout état de cause le juste milieu serait un exposé des éléments de la théorie de l'estimation et de test d'hypothèses combiné à la recherche systématique des procédures optimales.

L'objectif principal de cet ouvrage est d'exposer la réalité actuelle de la statistique mathématique sous la forme la plus accessible et la plus cohérente qui soit.

Cet ouvrage se compose de 5 chapitres et d'Annexes.

Le chapitre 1 traite des propriétés (essentiellement asymptotiques) des distributions empiriques, qui sont à la base de la statistique mathématique.

Les chapitres 2 et 3 développent respectivement la théorie des estimateurs et la théorie de test d'hypothèses statistiques. Les premières parties de chacun de ces chapitres sont consacrées à la description des éventuelles méthodes de résolution des problèmes posés et à la recherche des procédures optimales, les deuxièmes, à la construction des procédures asymptotiquement normales.

Le chapitre 5 qui expose la théorie des jeux dans les problèmes de statistique mathématique présente la même structure.

Le chapitre 4 étudie les problèmes relatifs à deux échantillons et plus.

Cet ouvrage comporte aussi 8 Annexes qui donnent les démonstrations de théorèmes, qui sortent du cadre de l'exposé principal par leur nature ou par leur complexité.

Les remarques bibliographiques qui sont loin d'être exhaustives permettent néanmoins de se faire une idée de l'émergence et de l'évolution des principaux domaines de la statistique mathématique. Les références renvoient de préférence aux monographies (car plus accessibles) qu'aux articles.

Parmi les nombreux ouvrages de statistique mathématique, nous en distinguerons quatre tant ils nous semblent complets et répondre à l'esprit actuel de cette matière : il s'agit des livres de H. Cramer [19], E. Lehmann [50], Sh. Zaks [91] et I. Ibraguimov & R. Khasminsky [42]. L'exposé de cet ouvrage a été le plus influencé par [42] (dont certaines idées sont utilisées dans les §§ 23, 24, 25, 27, 28, 29 du chapitre 2) et par [50]. Le reste est peu lié à la structure des ouvrages existants.

D'innombrables autres ouvrages ont fortement marqué la statistique mathématique (notamment ceux de D. Blackwell et Girshik [7], M. Kendall et A. Stuard [43], T. Ferguson [27], C. Rao [68], etc. Qu'on nous pardonne de ne pouvoir les citer tous) mais ils se distinguent fondamentalement de cette monographie tant par leur esprit que par leur contenu.

En plus des résultats et méthodes classiques, cet ouvrage propose des rubriques nouvelles qui facilitent l'exposé, des améliorations méthodologiques ainsi que des résultats nouveaux et des résultats publiés pour la première fois dans une monographie.

Voici brièvement décrits les grands traits de cet ouvrage.

Dans les §§ 1 et 2 du chapitre 1 on introduit les notions d'échantillon, de distribution empirique et on établit le théorème de Glivenko-Cantelli qui peut être considéré comme un fait fondamental sur lequel reposent les inférences statistiques.

Dans le § 3 on étudie deux types de statistiques qui englobent l'écrasante majorité des statistiques pratiquement intéressantes. Ces statistiques sont définies comme les valeurs $G(\mathbf{P}_n^*)$ de fonctionnelles G (satisfaisant certaines conditions) dépendant d'une distribution empirique $\mathbf{P}_n^*$. On démontre plus loin (§§ 7 et 8) les théorèmes limites relatifs à la distribution de ces statistiques. Ceci allège la suite de l'exposé et nous libère de la nécessité de reproduire pour chaque statistique pratiquement les mêmes raisonnements, des raisonnements qui de surcroît sont sans rapport avec le fond du problème.

Le § 5 regroupe des théorèmes auxiliaires (appelés ici « théorèmes de continuité ») sur la convergence des distributions et de leurs moments. Ceci allège aussi la suite de l'exposé.

Dans le § 6 (que l'on peut omettre en première lecture) on montre que la fonction empirique de répartition $F_n^*(t)$ est un processus poissonnien con-

ditionnel et on énonce le théorème de convergence du processus $\sqrt{n}(F_n^*(t) - F(t))$ vers un pont brownien (la démonstration de ce théorème est donnée dans l'Annexe I).

Dans le § 10 on introduit les distributions empiriques lissées qui permettent d'approcher non seulement une distribution mais aussi sa densité.

Dans le § 3 du chapitre 2 qui est consacré aux estimateurs de paramètres inconnus, on développe une approche unique de construction des estimateurs appelée « méthode de substitution ». Cette méthode nous suggère de chercher un estimateur θ^* du paramètre θ, représenté par une fonctionnelle $\theta = G(\mathbf{P})$ dépendant de la distribution $\mathbf{P}$ de l'échantillon, sous la forme $\theta^* = G(\mathbf{P}_n^*)$, où $\mathbf{P}_n^*$ est une distribution empirique. Tous les estimateurs « raisonnables » utilisés dans la pratique sont des estimateurs de substitution. L'optimalité d'un estimateur est atteinte par un choix convenable de la fonctionnelle G. Si une statistique $\theta^* = G(\mathbf{P}_n^*)$ est une statistique de type I ou II, les théorèmes du chapitre 1 nous permettent d'établir immédiatement la convergence et la normalité asymptotique de cet estimateur. Dans les §§ 4 et 5, cette approche est illustrée sur des estimateurs de la méthode des moments et de la méthode du minimum de la distance. On aurait pu envisager les estimateurs du maximum de vraisemblance du même point de vue (§ 6), mais leur étude directe permet d'établir des résultats plus profonds indispensables pour la suite.

Dans le chapitre 2 on développe deux approches pour la comparaison des estimateurs : l'approche *de la moyenne quadratique* (on compare $\mathsf{E}_\theta(\theta^* - \theta)^2$) et l'approche *asymptotique* (on compare les variances de la distribution limite de $\sqrt{n}(\theta^* - \theta)$ dans la classe des estimateurs asymptotiquement normaux). Dans le cas paramétrique ceci permet de distinguer 3 types d'estimateurs optimaux : les estimateurs efficaces dans les classes K_b de biais fixe, les estimateurs bayésiens et les estimateurs minimax. On utilise les mêmes principes pour dégager les classes des estimateurs *asymptotiquement* optimaux dans l'approche asymptotique. Les estimateurs efficaces sont construits à l'aide des méthodes traditionnelles suivantes : la première est qualitative et s'appuie sur le principe d'exhaustivité (§§ 12, 13 et 14) ; la deuxième est basée sur des relations qualitatives découlant de l'inégalité de Rao-Cramer (§ 16) ; la troisième repose sur des considérations d'invariance (§§ 17 et 19) qui permettent de restreindre les classes d'estimateurs envisagées. La recherche des estimateurs asymptotiquement optimaux et l'étude des propriétés asymptotiques de la fonction de vraisemblance font l'objet des §§ 20 à 30. Le § 20 contient une inégalité intégrale de type Rao-Cramer qui permet, en particulier, d'établir des critères simples pour qu'un estimateur soit asymptotiquement bayésien et minimax et de justifier le choix de la sous-classe $\tilde{K}_0$ à laquelle il faut limiter la recherche des estimateurs asymptotiquement efficaces. Ceci permet d'établir immédiatement par une

étude des propriétés asymptotiques des estimateurs du maximum de vraisemblance (§ 25) que ces estimateurs sont asymptotiquement bayésiens et minimax et asymptotiquement efficaces dans $\bar{K}_0$. Les §§ 21 à 24 sont accessoires. L'estimation par intervalles est traitée dans les §§ 31 et 32 et dans le § 8 du chapitre 3.

Le chapitre 3 est consacré au test d'hypothèses. Les §§ 1 et 2 traitent le cas d'un nombre fini d'hypothèses simples. On distingue (comme dans la théorie de l'estimation) trois types de tests optimaux : les tests les plus puissants dans des sous-classes, les tests bayésiens et les tests minimax. On établit les liens existant entre ces tests et on les détermine sous forme explicite. On se base sur le principe de Bayes (et non pas sur le lemme de Neyman-Pearson) ce qui à notre sens simplifie l'exposé et le rend plus limpide. Dans le § 3 on développe les approches asymptotiques de calcul des tests de deux hypothèses simples et on les compare. Dans le § 4 on considère la position générale du problème de test de deux hypothèses composées et l'on définit les classes de tests optimaux (uniformément les plus puissants, bayésiens et minimax). Le § 5 est consacré à la recherche des tests uniformément les plus puissants dans les cas où cela est possible. Dans les §§ 6 et 7, on se penche sur le même problème mais dans des classes de tests restreintes pour des raisons d'invariance et d'absence de biais. Comme dans les §§ 1 et 2, cette étude est conduite du point de vue bayésien. Dans le § 8, les résultats acquis sont appliqués à la construction des régions de confiance les plus exactes. Dans le § 9 on étudie les tests bayésiens et les tests minimax. Les §§ 10 et 13 traitent du test du rapport de vraisemblance. Ce test est uniformément le plus puissant dans de nombreux cas particuliers et est asymptotiquement bayésien sous des conditions assez larges. Dans les §§ 15, 16 et 17 on poursuit l'étude des propriétés d'optimalité asymptotique du test du rapport de vraisemblance. Le § 11 établit l'optimalité de ce test dans les problèmes d'analyse de variance. Les § 14 et 15 sont consacrés à la recherche de tests asymptotiquement optimaux d'hypothèses proches et à leur détermination sous forme explicite pour les principaux problèmes de statistique.

Les trois premiers chapitres ont pour caractéristique essentielle de ne traiter que des problèmes de statistique portant sur un seui échantillon.

Le chapitre 4, comme déjà signalé, est consacré aux problèmes faisant intervenir deux échantillons et plus, notamment les problèmes d'homogénéité (totale ou partielle, §§ 1 et 2), problèmes de régression (§ 3) et les problèmes d'analyse de variance (§ 4). Les résultats du chapitre 3 sont appliqués aux problèmes d'homogénéité dans le cas paramétrique pour construire des tests asymptotiquement optimaux sous la condition que les hypothèses alternatives soient proches de l'hypothèse de base d'homogénéité. Les résultats des chapitres 2 et 3 sont utilisés dans les problèmes de régression (aussi bien de régression linéaire que de régression en des fonctions arbitraires) pour construire des estimateurs efficaces pour les paramètres inconnus et

des tests pour éprouver les hypothèses de base. On étudie également des problèmes d'analyse discriminante.

Le chapitre 5 traite de l'application de la théorie des jeux à la résolution de problèmes de statistique mathématique. Ce chapitre contribue à élaborer une vision globale de la statistique mathématique et permet de généraliser de nombreux résultats des chapitres 2 et 3. Dans le § 2 on développe les notions fondamentales et les résultats de la théorie des jeux « ordinaire » (on ne considère que les jeux à deux personnes). On établit en particulier des liens entre les principaux types de stratégies optimales : les stratégies bayésiennes, minimax et uniformément les meilleures dans des sous-classes. Dans le § 3 on étudie les jeux statistiques. Dans le § 4, on énonce et on prouve le principe de Bayes qui permet de ramener la recherche d'une décision bayésienne à un problème plus simple de construction d'une stratégie bayésienne pour un jeu ordinaire à deux joueurs. Dans le § 5, on discute les principes d'exhaustivité, d'absence de biais et d'invariance pour la construction des décisions uniformément les meilleures dans les sous-classes correspondantes. Les §§ 6, 7 et 8 sont consacrés à la recherche de décisions asymptotiquement optimales. Dans le § 6 on se penche sur les estimateurs asymptotiquement optimaux des paramètres pour une fonction de perte quelconque (et pas seulement quadratique). On réussit dans ce cas à établir des résultats proches de ceux du chapitre 2 sur l'optimalité asymptotique des estimateurs du maximum de vraisemblance. Dans les §§ 7 et 8 on traite les tests asymptotiquement optimaux pour une fonction de perte quelconque. Dans le § 7, on prouve que le test du rapport de vraisemblance est asymptotiquement bayésien ; dans le § 8, on établit le critère limite d'optimalité des tests d'hypothèses voisines (généralisation des résultats des §§ 14, 15 du chapitre 3 au cas d'une fonction de perte quelconque).

De toutes les Annexes on distinguera l'Annexe VIII dans laquelle sont prouvés deux théorèmes fondamentaux de la théorie des jeux statistiques, dont la lecture nécessite des connaissances mathématiques poussées.

Cet ouvrage se fixe de multiples objectifs. Certes par son contenu il est plus proche du niveau de la maîtrise, mais les mesures prises pour en faciliter la première lecture le mettent à la portée des étudiants. Les paragraphes « ultra-compliqués » ou « plus avancés », qui sont signalés par un astérisque, doivent être sautés en première lecture au même titre d'ailleurs que les passages en petits caractères. D'autre part les cas techniquement plus compliqués portant sur un paramètre vectoriel sont presque partout traités dans des rubriques et des paragraphes à part sur lesquels on peut aussi glisser.

Les professeurs qui ont une connaissance partielle de cette matière peuvent sélectionner un ensemble de paragraphes pour concocter un cours de statistique mathématique étalé sur un semestre. Entre autres variantes nous leur conseillons les §§ 1, 3 et 5 du chapitre 1 ; les §§ 2, 3, 4, 6 à 12, 14, 16 (21, 23, 24, 25), 31 et 32 du chapitre 2 et les §§ 1, 2, 4, 5, 12 (13, 16) du chapi-

tre 3. Les paragraphes placés entre parenthèses sont consacrés aux procédures asymptotiquement optimales. On peut les alléger au maximum ou tout simplement les omettre en fonction du niveau de préparation des étudiants.

La lecture de cet ouvrage suppose connus les éléments fondamentaux de la théorie des probabilités tels qu'ils sont présentés dans le manuel [11] du même auteur. Au contraire des autres, les références à cet ouvrage apparaissent dans les passages qui sont supposés être connus du lecteur, et essentiellement à titre de rappel.

Les paragraphes, théorèmes, lemmes, exemples, etc., sont numérotés de façon autonome. Pour faciliter la lecture, on se refère à un théorème ou autre de façon différente selon la place qu'il occupe par rapport au passage étudié. Ainsi

cf. théorème 1 ou formule (12) = théorème 1 ou formule (12) du paragraphe étudié ;

cf. théorème 13.1 ou formule (13.12) = théorème 1 du § 13 ou formule (12) du § 13 du chapitre étudié ;

cf. théorème 2.13.1 ou formule (2.13.12) = théorème 1 § 13 chapitre 2 ou formule (12) § 13 chapitre 2.

Idem pour les paragraphes :

cf. § 13 renvoie au § 13 du chapitre étudié ;

cf. § 2.13 renvoie au § 13 du chapitre 2.

Le signe ◄ marque la fin d'une démonstration.

La composition de cet ouvrage s'est faite en plusieurs étapes assez laborieuses et doit à beaucoup de personnes :

à I. Borissov qui a apporté une aide inappréciable à la préparation et la correction du manuscrit ;

à K. Borovkov qui m'a prodigué des conseils utiles et dont les remarques ont contribué à « purifier » le texte final ;

à A. Sakhanenko qui a bien voulu lire le manuscrit à ma demande et dont les suggestions ont été utilisées notamment dans les démonstrations des §§ 16, 21, 23, 29 du chapitre 2, les §§ 13 à 15 du chapitre 3, les Annexes II, V et VIII (voir également la notice bibliographique) ;

à D. Tchibissov dont les remarques précieuses ont contribué à améliorer le contenu ;

à V. Yourinski et A. Novikov qui ont lu le manuscrit et dont les suggestions m'ont été d'une grande utilité.

A ces personnes et à toutes celles qui m'ont apporté leur aide sous quelque forme que ce soit je voudrais exprimer ici ma profonde et sincère gratitude.

A. Borovkov

Décembre 1985

INTRODUCTION

Cet ouvrage expose les fondements de la statistique mathématique appelée parfois tout simplement *statistique*. Mais cette abréviation ne doit pas prêter à équivoque dans la mesure où le terme de statistique recouvre généralement un autre sens.

Qu'est-ce que la statistique mathématique ? Il existe plusieurs définitions descriptives qui reflètent à des degrés différents le contenu de cette discipline mathématique. L'une des plus simples et des plus vagues repose sur une comparaison liée à la notion d'échantillon d'une population générale et au problème de la distribution hypergéométrique qui est généralement traitée au début de tout cours de théorie des probabilités. Ce *problème direct* de la théorie des probabilités consiste à étudier la distribution de la composition d'un échantillon aléatoire au vu de la composition de la population générale. Mais le *problème inverse* qui consiste à reconstruire la population générale au vu d'un échantillon se pose fréquemment. Ce sont précisément ces problèmes qui font à proprement parler l'objet de la statistique mathématique.

Précisons cette comparaison : en théorie des probabilités on connaît la nature d'un phénomène et on cherche à comprendre le comportement (la distribution) des caractéristiques observées dans des expériences. En statistique mathématique, c'est l'inverse : on part des données expérimentales (généralement des observations des variables aléatoires) et on demande d'émettre un jugement ou de prendre une décision sur la nature du phénomène étudié. Nous touchons ainsi à l'une des plus importantes sphères de l'activité humaine : la connaissance. L'idée que le « critère de la vérité est la pratique » trouve sa pleine justification en statistique mathématique dans la mesure où cette science précisément étudie les méthodes (dans le cadre de modèles mathématiques exacts) qui nous permettent de dire si les résultats de l'expérience confirment ou infirment l'hypothèse avancée sur la nature du phénomène.

Fait important, comme en théorie des probabilités, on s'intéresse non pas aux expériences à issues déterministes, mais aux expériences donnant

lieu à des événements aléatoires. Le rôle de ces problèmes ne cesse de s'amplifier avec les progrès de la science, car avec l'accroissement de la précision des expériences il devient de plus en plus difficile d'éviter le « facteur aléatoire » dû à toute sorte de perturbations et aux possibilités restreintes des instruments de mesure et de calcul.

La statistique mathématique est une partie de la théorie des probabilités en ce sens que chaque problème de la première est en fait un problème (parfois assez original) de la seconde. Mais la statistique mathématique est aussi une science autonome qui peut être considérée comme une science sur le comportement inductif de l'homme (et pas seulement de l'homme) lorsqu'il doit sur la base de son expérience non déterministe prendre des décisions avec des pertes minimales *).

La statistique mathématique s'appelle aussi théorie des décisions statistiques, puisqu'on peut la caractériser comme la science des décisions optimales (ces deux termes sont à préciser) basées sur des données statistiques (empiriques). Les positions exactes des problèmes seront données dans le texte de cet ouvrage. Nous nous bornerons ici à trois exemples de problèmes de statistique les plus simples et les plus typiques.

EXEMPLE 1. L'un des principaux paramètres caractérisant la qualité d'un article est la durée de service. Mais cette durée est en principe aléatoire et impossible à déterminer à l'avance. L'expérience montre que si le processus de production est dans un certain sens homogène, les durées de service $\xi_1, \xi_2, \ldots$ respectivement du 1-ier, 2-ième, etc. article peuvent être traitées comme des variables aléatoires indépendantes équidistribuées. Il est alors naturel d'identifier le paramètre durée de service au nombre $\theta = \mathsf{E}\xi_i$. Un problème classique consiste à déterminer la valeur de θ. Pour ce faire, on prend n articles et on les contrôle. Soient $x_1, x_1, \ldots, x_n$ les durées de services des articles contrôlés. On sait que

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i \underset{\text{p.s.}}{\to} 0 \text{ lorsque } n \to \infty.$$

Il est intuitif que le nombre $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ soit proche de θ pour n assez grand et permette dans une certaine mesure de répondre à la question posée. Ceci étant, il est évident que le nombre d'observations n doit être le plus

*) Pour plus de détails voir [59].

petit possible et l'estimation de θ la meilleure possible (une valeur trop forte ou trop faible du paramètre θ conduirait à des pertes matérielles).

EXEMPLE 2. Un radar sonde une partie de l'espace aux instants $t_1, t_2, \ldots$ $\ldots, t_n$ afin de détecter un certain objet. Désignons par $x_1, \ldots, x_n$ les valeurs des signaux réfléchis captés par le radar. Si l'objet cherché ne se trouve pas dans cette partie de l'espace, les valeurs x_i peuvent être traitées comme des variables aléatoires indépendantes distribuées comme une variable aléatoire ξ dont la nature dépend du caractère des diverses perturbations. Si tout au long des observations l'objet se trouve dans le champ de vision, les x_i contiendront le signal « utile » a avec les perturbations et seront alors distribuées comme $\xi + a$. Donc, si dans le premier cas la fonction de répartition des observations x_i était $F(x)$, dans le second elle sera de la forme $F(x - a)$. On demande de dire au vu des observations $x_1, \ldots, x_n$ si l'objet se trouve ou non dans la zone scrutée.

Dans ce problème il apparaît possible d'indiquer dans un certain sens la « règle de décision optimale » qui donnera la solution du problème avec des erreurs minimales. Ce problème peut être compliqué de la manière suivante. L'objet n'apparaît dans la zone visée qu'à partir d'une observation de numéro inconnu θ. On demande de déterminer avec le plus de précision l'instant θ d'apparition de l'objet. Ceci est le « problème de panne » qui admet une foule d'interprétations importantes dans les applications.

EXEMPLE 3. Une expérience est effectuée d'abord n_1 fois dans des conditions A et ensuite n_2 fois dans des conditions B. Soient $x_1, \ldots, x_{n_1}$ et $y_1, \ldots, y_{n_2}$ les résultats de ces expériences respectivement dans les conditions A et B. On demande de dire si le changement de conditions se répercute sur les résultats. En d'autres termes, si l'on désigne par $\mathbf{P}_A$ la distribution de x_i, $1 \leqslant i \leqslant n_1$, et par $\mathbf{P}_B$, la distribution de y_i, $1 \leqslant i \leqslant n_2$, le problème consiste à dire si la relation $\mathbf{P}_A = \mathbf{P}_B$ est remplie ou non.

Si par exemple l'on étudie l'influence d'un produit sur la croissance, disons, de plantes ou d'animaux, on procède à deux séries d'expériences parallèles (avec et sans le produit) et l'on compare les résultats obtenus.

On est souvent confronté à des problèmes plus compliqués où la même question est posée pour plusieurs séries d'observations réalisées dans des conditions différentes. Si les résultats des observations dépendent des conditions, il est nécessaire de vérifier le caractère de cette dépendance (ceci est le problème de regression).

L'exemple 3 et les problèmes plus compliqués cités font partie de la classe des problèmes *à deux échantillons* et *plus*. De tels problèmes seront envisagés dans le chapitre 4.

On pourrait prolonger la liste des exemples de problèmes typiques de statistique, problèmes qui diffèrent autant par leur contenu que par leur

complexité. Mais ces problèmes ont en commun les deux aspects suivants :

1. Les distributions des résultats des observations sont inconnues.

2. Dans chacun de ces problèmes, il faut prendre au vu des résultats des observations une décision sur la distribution de ces observations (d'où le nom de « Théorie des décisions statistiques » mentionné plus haut).

En vertu de ces remarques, le fait suivant revêt une signification fondamentale pour la suite et en particulier pour la résolution des problèmes cités dans les exemples. Il apparaît qu'au vu des observations $x_1, \ldots, x_n$ d'une variable aléatoire ξ on peut reconstituer la distribution inconnue $\mathbf{P}$ de ξ avec la précision que l'on veut pour les grands n. Ceci vaut également pour toute fonctionnelle $\theta = \theta(\mathbf{P})$.

A ce fait qui repose à la base de la statistique mathématique et aux positions plus exactes des problèmes sera consacré le chapitre 1.

CHAPITRE PREMIER

ÉCHANTILLON. DISTRIBUTION EMPIRIQUE. PROPRIÉTÉS ASYMPTOTIQUES DES STATISTIQUES

Dans les §§ 1 à 4 on introduit les notions d'échantillon et de distribution empirique et on étudie leurs propriétés élémentaires, essentiellement les propriétés asymptotiques à la base de la statistique mathématique.

Le § 5 est consacré aux théorèmes de continuité (de convergence des distributions de fonctions de suites de variables aléatoires) qui seront utilisés tout au long de cet ouvrage.

Les §§ 6 à 10 traitent des plus fines propriétés asymptotiques des distributions empiriques, et des distributions limites des principaux types de statistiques.

§ 1. Notion d'échantillon

Toute étude statistique repose sur un ensemble d'observations. Dans les cas les plus simples ce sont les valeurs empiriques (obtenues dans le cadre d'une expérience) d'une variable aléatoire ξ. Nous avons signalé que dans les problèmes de statistique, la distribution **P** de cette variable aléatoire est au moins partiellement inconnue.

Plus exactement, soit G une expérience rattachée à une variable aléatoire ξ. Formellement, nous devons construire pour cette expérience un modèle mathématique mettant en jeu un espace probabilisé $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mathbf{P})$, et définir de façon convenable sur cet espace une fonction mesurable appelée *variable aléatoire* ξ (cf. [11]). Sans nuire à la généralité, on peut admettre que l'espace $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mathbf{P})$, est un *espace des échantillons* (cf. [11]), autrement dit admettre que $\mathscr{X}$ est l'espace des valeurs de $\xi(x) = x$. Dans ce cas, **P** peut être appelée *distribution* (ou *loi de probabilité*) de ξ. Si ξ est une variable aléatoire numérique, $\mathscr{X}$ est la droite numérique R ; si ξ est un vecteur, $\mathscr{X}$ est un espace R^m à $m > 1$ dimensions. Dans la suite, nous n'envisagerons que ces deux cas, c'est-à-dire que $\mathscr{X}$ sera la droite R pour $m = 1$ ou l'espace R^m pour $m > 1$. Pour $\mathfrak{B}_{\mathscr{X}}$ on prend souvent la tribu des boréliens de $\mathscr{X}$*).

Si l'on sait *a priori* que **P** est concentrée sur une partie $B \in \mathfrak{B}_{\mathscr{X}}$ de l'espace $\mathscr{X}$, on aura intérêt à assimiler $\mathscr{X}$ à B et $\mathfrak{B}_{\mathscr{X}}$ à sa restriction à B.

---

*) De nombreuses sections de cet ouvrage sont valables pour une situation plus générale où $\mathscr{X}$ est un espace métrique quelconque et $\mathfrak{B}_{\mathscr{X}}$ la tribu de ses boréliens, c'est-à-dire la tribu de ses ouverts.

Considérons n répétitions indépendantes de l'expérience G (cf. [11]) et désignons par $x_1, \ldots, x_n$ l'ensemble de valeurs observées. Le vecteur

$$X_n = (x_1, \ldots, x_n)$$

s'appelle *échantillon de taille n prélevé dans une population de distribution* **P**. Pour abréger on dira aussi « un échantillon issu de la distribution **P** » ou encore « échantillon de distribution **P** ».

Pour noter que « X_n est un échantillon de distribution **P** », on se servira du symbole :

$$X_n \in \mathbf{P}. \tag{1}$$

On utilisera cette notation pour les autres variables aléatoires. Ainsi,

$$\xi \in \mathbf{P} \tag{2}$$

voudra dire que ξ admet la distribution **P**. Cet usage du symbole $\in$ est conforme à (1), puisque (1) est définie pour tout n et en particulier pour $n = 1$.

Si ξ et η sont des variables aléatoires (en général définies sur des espaces différents) de même distribution, on dira qu'elles sont *parentes* ou encore *équidistribuées* et on notera ce fait par $\xi \underset{d}{=} \eta$. Si X_n et Y_n sont des échantillons de même taille, de distribution **P**, on dira aussi qu'ils sont *parents* et on écrira : $X_n \underset{d}{=} Y_n$.

Aux seconds membres de (1) et (2) peut parfois figurer la fonction de répartition de **P**. Si $F(x) = \mathbf{P}(]-\infty, x[)$, la notation

$$X_n \in F$$

sera indentique à (1).

La notion d'« échantillon d'une population générale » se rencontre aussi dans les modèles probabilistes élémentaires liés au tirage de boules d'une urne dans la définition classique de la probabilité (cf. [11], § 2 du chap. 1). A noter que la définition donnée ici de l'échantillon est de même nature, voire même confondue avec celle du [11]. Si les x_i (ou la variable aléatoire ξ) ne peuvent prendre que s valeurs $a_1, \ldots, a_s$ et que les probabilités de ces valeurs soient rationnelles, c'est-à-dire

$$\mathbf{P}(\xi = a_j) = \frac{N_j}{N}, \quad \sum_{j=1}^{s} N_j = N,$$

l'échantillon X_n peut être traité comme le résultat d'un « tirage avec remise » (au sens du chapitre 1, [11]) dans une urne contenant N boules dont N_1 sont marquées a_1, N_2 sont marquées a_2 et ainsi de suite.

Donc, l'être mathématique $X = X_n$ (nous omettrons souvent l'indice n) n'est autre qu'une variable aléatoire $(x_1, \ldots, x_n)$ à valeurs dans un espace

« n-dimensionnel » $\mathscr{X}^n = \mathscr{X} \times \mathscr{X} \times \ldots \times \mathscr{X}$, et de distribution définie pour $B = B_1 \times B_2 \times \ldots \times B_n$, $B_j \in \mathfrak{B}_{\mathscr{X}}$, par les égalités

$$\mathbf{P}(X \in B) = \mathbf{P}(\mathrm{x}_1 \in B_1, \ldots, \mathrm{x}_n \in B_n) = \prod_{i=1}^{n} \mathbf{P}(\mathrm{x}_i \in B_i). \quad (3)$$

En d'autres termes, la distribution $\mathbf{P}$ définie sur $\mathscr{X}^n$ est le produit direct de n distributions « unidimensionnelles » données.

S'agissant des notations de $\mathbf{P}$ et des autres, on adoptera les conventions suivantes que l'on a déjà utilisées dans (3), sans risque d'ambiguïté.

1. On se servira du même symbole (en l'occurrence $\mathbf{P}$) pour les distributions sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ et pour le produit direct de ces distributions sur $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ (cf. (3)), où $\mathfrak{B}^n_{\mathscr{X}}$ est la tribu des boréliens de $\mathscr{X}^n$. La seule différence viendra de l'argument de la fonction $\mathbf{P}$.

2. La probabilité qu'une quantité X tombe dans un ensemble B, par exemple de $\mathfrak{B}^n_{\mathscr{X}}$, nous la noterons soit pas $\mathbf{P}(B)$, soit par $\mathbf{P}(X \in B)$ selon les besoins. Ces notations sont identiques, puisque $\mathscr{X}^n$ est l'espace des échantillons X.

3. Enfin, le symbole $\mathbf{P}$ désignera la notion générale de probabilité (c'est-à-dire la probabilité rattachée à d'autres variables aléatoires sans spécifier l'espace probabilisé).

En vertu de (3), on peut traiter l'échantillon X comme un événement élémentaire dans un espace probabilisé des échantillons $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}}, \mathbf{P})$ (cf. [11], chap. 3, § 2). Signalons que X peut être interprété, tantôt comme une variable aléatoire, tantôt comme un vecteur dont les coordonnées sont les valeurs numériques obtenues au cours d'une expérience. L'usage montre que cette double interprétation est convenable et ne prête pas à équivoque malgré l'existence simultanée de notations de la forme $\mathbf{P}(\mathrm{x}_1 < t) = F(t)$, $\mathrm{x}_1 = 0{,}74$, $\mathrm{x}_2 = 0{,}83$, etc.

Signalons également que les composantes x_i de l'échantillon X seront désignées par des lettres minuscules « droites » x et les variables par des lettres italiques. Le vecteur $(x_1, \ldots, x_n) \in \mathscr{X}^n$, $x_i \in \mathscr{X}$, sera représenté par une lettre semi-grasse $\boldsymbol{x} = (x_1, \ldots, x_n)$.

L'échantillon est le principal élément liminaire dans les problèmes de statistique mathématique. Ses composantes $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$ ne sont pas toujours indépendantes. Dans la suite nous n'excluerons pas cette éventualité. Mais pour ne pas poser de conditions supplémentaires, nous admettrons en cas d'observations dépendantes que nous avons affaire à un échantillon de taille $n = 1$ et que les observations sont les coordonnées d'un vecteur x_1 (l'espace $\mathscr{X}$ n'est-il pas de nature arbitraire !)

Nous aurons souvent à considérer des échantillons X_n de taille n illimitée. Dans ces cas, il sera commode de postuler qu'est donné un échantillon

$X_\infty = (x_1, x_2, \ldots)$ de taille infinie dont $X = X_n$ serait l'ensemble de ses n premières composantes. Par échantillon X_∞ de taille infinie, on entendra un élément de l'espace probabilisé $(\mathscr{X}^\infty, \mathfrak{B}^\infty_{\mathscr{X}}; \mathbf{P})$, où $\mathscr{X}^\infty$ est l'espace des suites $(x_1, x_2, \ldots)$, $\mathfrak{B}^\infty_{\mathscr{X}}$ la tribu des ensembles $\bigcap_{i \leq N} \{x_i \in B_i\}$, $B_i \in \mathfrak{B}_{\mathscr{X}}$; $N = 1, 2, \ldots$ et la distribution $\mathbf{P}$ possède la propriété (3). Le théorème de Kolmogorov ([11]) affirme qu'une telle distribution existe toujours. Donc, l'hypothèse qu'il existe un échantillon X_∞ ne restreint en aucun cas la généralité.

La suite (l'échantillon) infinie X_∞ peut être traitée comme un événement élémentaire (cf. [11]) dans un cadre probabiliste.

Dans les cas où nous aurons besoin de comprendre X_n comme un sous-vecteur de X_∞, on écrira

$$X_n = [X_\infty]_n,$$

où $[\cdot]_n$ est l'opérateur de projection de $\mathscr{X}^\infty$ sur $\mathscr{X}^n$ qui se définit de manière évidente. Conformément à ce qui précède, la notation

$$X_\infty \in \mathbf{P}$$

exprimera que X_∞ est un échantillon de taille infinie de distribution $\mathbf{P}$.

S'il est indispensable de mentionner expressément que l'on étudie une distribution sur $(\mathscr{X}^\infty, \mathfrak{B}^\infty_{\mathscr{X}})$ (resp. sur $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ pour $n < \infty$), et non sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$, on se servira de la notation $\mathbf{P}^\infty$ (resp. $\mathbf{P}^n$). L'utilisation systématique des indices supérieurs ∞ et n alourdirait considérablement les notations.

§ 2. Distribution empirique (en dimension un)

Soit donné un échantillon $X = (x_1, \ldots, x_n) \in \mathbf{P}$, $x_i \in \mathscr{X} = R$. Considérons la droite réelle R munie de la tribu de ses boréliens, et une distribution discrète $\mathbf{P}^*_n$ sur $(R, \mathfrak{B})$, concentrée aux points $x_1, \ldots, x_n$, et telle que la probabilité de x_i est égale à $1/n$. Autrement dit, pour tout $B \in \mathfrak{B}$, on a par définition

$$\mathbf{P}^*_n(B) = \frac{\nu(B)}{n}, \tag{1}$$

où $\nu(B)$ est le nombre d'éléments de l'échantillon X contenus dans l'ensemble B. La distribution $\mathbf{P}^*_n$ s'appelle *distribution empirique* construite au vu de l'échantillon X (ou associée à l'échantillon X). On peut la représenter encore sous la forme suivante. Soit $\mathbf{I}_x(B)$ une distribution concentrée en un point x :

$$\mathbf{I}_x(B) = \begin{cases} 1, & x \in B, \\ 0, & x \notin B; \end{cases}$$

il est évident que $\nu(B) = \sum_{i=1}^{n} \mathbf{I}_{x_i}(B)$ et

$$\mathbf{P}_n^*(B) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{x_i}(B). \tag{2}$$

Il est clair que pour tout borélien B, la distribution $\mathbf{P}_n^*(B)$ traitée comme une fonction d'échantillon est une variable aléatoire. Nous avons donc affaire à une fonction d'ensemble aléatoire ou à une distribution aléatoire.

Supposons maintenant que $X_\infty \in \mathbf{P}$, $X_n = [X_\infty]_n$ et $n \to \infty$. Nous obtenons alors une suite de distributions empiriques $\mathbf{P}_n^*$. Le fait remarquable est que cette suite se rapproche indéfiniment de la distribution primitive $\mathbf{P}$ de la variable aléatoire observée. Ce fait est capital pour la suite de l'exposé, car il indique que la distribution inconnue $\mathbf{P}$ peut être reconstituée, avec la précision que l'on veut, sur le vu d'un échantillon de taille assez élevée.

THÉORÈME 1. *Soient* $B \in \mathfrak{B}$ *et* $X_n = [X_\infty]_n \in \mathbf{P}$. *Alors pour* $n \to \infty$

$$\mathbf{P}_n^*(B) \underset{\text{p.s.}}{\to} \mathbf{P}(B).$$

La convergence presque sûre (c'est-à-dire avec la probabilité 1) est entendue pour la distribution $\mathbf{P} = \mathbf{P}^\infty$ sur $(R^\infty, \mathfrak{B}^\infty, \mathbf{P})$. Nous avons introduit l'hypothèse $X_n = [X_\infty]_n$ pour définir les variables aléatoires $\mathbf{P}_n^*(B)$ sur un même espace probabilisé.

DÉMONSTRATION. Tournons-nous vers la définition (2) pour remarquer que $\mathbf{I}_{x_i}(B)$ sont des variables aléatoires indépendantes équidistribuées : $\mathsf{E}\mathbf{I}_{x_i}(B) = \mathbf{P}(\mathbf{I}_{x_i}(B) = 1) = \mathbf{P}(x_i \in B) = \mathbf{P}(B)$. Puisque $\mathbf{P}_n^*(B)$ est la moyenne arithmétique de ces variables, il reste à appliquer la loi forte des grands nombres. ◄

Le théorème 1 établit la convergence de $\mathbf{P}_n^*(B)$ et $\mathbf{P}(B)$ en chaque « point » B. On a toutefois une proposition plus forte qui dit que cette convergence est dans un certain sens uniforme par rapport à B.

Désignons par $\mathfrak{J}$ la famille des ensembles B qui sont des intervalles semi-ouverts de la forme $[a, b[$ à extrémités finies ou infinies et supposons encore que $X_n = [X_\infty]_n$.

THÉORÈME 2 (Glivenko-Cantelli).

$$\sup_{B \in \mathfrak{J}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{p.s.}}{\to} 0.$$

En réalité, aux noms de Glivenko et Cantelli est rattachée une proposition légèrement différente relative à la notion importante de *fonction de répartition empirique*. Par définition, c'est la fonction de répartition correspondant à $\mathbf{P}_n^*$. En d'autres termes, on appelle *fonction de répartition empiri-*

que $F_n^*(x)$ la fonction

$$F_n^*(x) = \mathbf{P}_n^*(]-\infty, x[).$$

La quantité $nF_n^*(x)$ est égale au nombre d'éléments de l'échantillon inférieurs strictement à x. Pour construire $F_n^*(x)$ on se sert de la procédure suivante. On range les éléments $x_1, \ldots, x_n$ de l'échantillon par ordre de grandeur croissante :

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)} .$$

La suite obtenue s'appelle *échantillon ordonné* ou *série variationnelle.* On peut alors poser

$$F_n^*(x) = \frac{k}{n} \quad \text{pour} \quad x \in]x_{(k)}, x_{(k+1)}],$$

où k parcourt les valeurs de 0 à n, $x_{(0)} = -\infty$, $x_{(n+1)} = \infty$. Il est évident que $F_n^*(x)$ est une fonction en escalier présentant des sauts de $1/n$ aux points x_i si les x_i sont distincts.

Supposons que $F(x) = \mathbf{P}(]-\infty, x[)$ est la fonction de répartition de ξ (ou ce qui revient au même de x_1) et que $X_n = [X_\infty]_n$. Le théorème de Glivenko-Cantelli s'énonce comme suit :

THÉORÈME 2A. *Pour* $n \to \infty$

$$\sup_x |F_n^*(x) - F(x)| \underset{\text{p.s.}}{\to} 0.$$

On omettra l'indice n de F_n^* et on écrira simplement F^*.

DÉMONSTRATION du théorème 2A. Supposons tout d'abord pour simplifier que la fonction F est continue. Soit $\epsilon > 0$ un nombre aussi petit que l'on veut tel que $N = 1/\epsilon$ soit entier. Puisque F est continue, on peut exhiber des entiers $z_0 = -\infty, z_1, \ldots, z_{N-1}, z_N = \infty$, tels que

$$F(z_0) = 0,\ F(z_1) = \epsilon = 1/N, \ldots, F(z_k) = k\epsilon = k/N, \ldots, F(z_N) = 1.$$

Pour $z \in [z_k, z_{k+1}[$ on a les relations

$$\begin{aligned} F^*(z) - F(z) &\leqslant F^*(z_{k+1}) - F(z_k) = F^*(z_{k+1}) - F(z_{k+1}) + \epsilon, \\ F^*(z) - F(z) &\geqslant F^*(z_k) - F(z_{k+1}) = F^*(z_k) - F(z_k) - \epsilon. \end{aligned} \tag{3}$$

Appelons A_k l'ensemble des événements élémentaires $\omega = X_\infty$ sur lesquels $F^*(z_k) \to F(z_k)$. Le théorème 1 nous dit que $\mathbf{P}(A_k) = 1$. Donc, pour chaque $\omega \in A = \bigcap_{k=0}^{N} A_k$, il existe un $n(\omega)$ tel que pour tous les $n \geqslant n(\omega)$ l'on ait

$$|F^*(z_k) - F(z_k)| < \epsilon, \quad k = 0, 1, \ldots, N. \tag{4}$$

Jointes à (3) ces inégalités entraînent

$$\sup_z |F^*(z) - F(z)| < 2\epsilon. \tag{5}$$

Cette relation a donc lieu pour tout $\epsilon > 0$, tout $\omega \in A$ et tout $n \geqslant n(\omega)$ assez grand. Le théorème est prouvé pour la fonction continue F, puisque $\mathbf{P}(A) = 1$.

La démonstration est en tous points identique pour une fonction arbitraire $F(x)$. Il faut seulement se servir du fait suivant : pour toute fonction $F(x)$ il existe un nombre fini de points $-\infty = z_0 < z_1 < \ldots < z_{N-1} < z_N = \infty$ tels que

$$F(z_{k+1}) - F(z_k + 0) \leqslant \epsilon, \quad k = 0, 1, \ldots, N-1, \tag{6}$$

(pour fixer les idées on peut admettre que l'ensemble $\{z_j\}$ contient tous les points en lesquels F subit un saut supérieur par exemple à $\epsilon/2$). De façon exactement analogue à (3), on obtient pour $z \in]z_k, z_{k+1}]$

$$\begin{aligned} F^*(z) - F(z) &\leqslant F^*(z_{k+1}) - F(z_{k+1}) + \epsilon, \\ F^*(z) - F(z) &\geqslant F^*(z_k + 0) - F(z_k + 0) - \epsilon. \end{aligned} \tag{7}$$

Aux ensembles A_k qui sont définis comme précédemment, ajoutons les ensembles A_k^+, $k = 0, 1, \ldots, N$, sur lesquels $F^*(z_k + 0) \to F(z_k + 0)$. Le théorème 1 nous dit que $\mathbf{P}(A_k) = \mathbf{P}(A_k^+) = 1$, de sorte que sur l'ensemble $A = \bigcap_{k=0}^{N} A_k A_k^+$ tel que $\mathbf{P}(A) = 1$ on a l'inégalité (4) ainsi que les inégalités

$$|F^*(z_k + 0) - F(z_k + 0)| < \epsilon, \quad k = 0, 1, \ldots, N,$$

pour $n \geqslant n(\omega)$ assez grand. Combinées à (7) ces inégalités entraînent (5). ◄

Le théorème 2A est un cas particulier du théorème 2, puisque l'ensemble $]-\infty, x[$ appartient à $\mathfrak{J}$; par ailleurs, on peut déduire facilement le théorème 2 du théorème 2A, car pour $B = [a, b[$

$$|\mathbf{P}_n^*(B) - \mathbf{P}(B)| \leqslant |F_n^*(b) - F(b)| + |F_n^*(a) - F(a)|,$$

si bien que

$$\sup_{B \in \mathfrak{J}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \leqslant \sup_{a,\, b}[|F_n^*(b) - F(b)| + |F_n^*(a) - F(a)|] \underset{\text{p.s.}}{\to} 0.$$

REMARQUE 1. Il est immédiat de voir que de tels raisonnements nous permettent de prendre pour famille d'ensembles $\mathfrak{J}$ dans le théorème 2 des familles d'intervalles ouverts $]a, b[$, d'intervalles fermés $[a, b]$ et de réunions d'un nombre fini ($\leqslant$ à un certain N) de ces intervalles.

Par ailleurs, si pour $\mathfrak{J}$ on prend une classe d'ensembles assez riche, le théorème 2 est mis en défaut. Si par exemple $\mathfrak{J}$ contient la réunion de tout

nombre fini d'intervalles, alors l'ensemble $B_n = \bigcup_{k=1}^{n}]x_k - 1/n^2, x_k + 1/n^2[\in \mathfrak{J}$, $\mathbf{P}_n^*(B_n) = 1$ et pour la distribution $\mathbf{P}$ qui est uniforme sur $[0, 1]$, on a $\mathbf{P}(B_n) \leqslant 2/n$, de sorte que

$$\sup_{B \in \mathfrak{J}} | \mathbf{P}_n^*(B) - \mathbf{P}(B) | \geqslant \mathbf{P}_n^*(B_n) - \mathbf{P}(B_n) \to 1.$$

Signalons en conclusion de ce paragraphe que la représentation (2) permet d'obtenir relativement au comportement asymptotique de $\mathbf{P}_n^*$ des théorèmes plus précis que ceux de type Glivenko-Cantelli (ces résultats seront exhibés aux §§ 4, 6). Pour illustrer les possibilités qui nous sont offertes ici, rappelons que l'expression $\sum_{i=1}^{n} \mathbf{I}_{x_i}(B)$ de (2) est une somme de variables aléatoires indépendantes équidistribuées dans le schéma de Bernoulli

$$\mathsf{E}\mathbf{I}_{x_i}(B) = \mathbf{P}(\mathbf{I}_{x_i}(B) = 1) = \mathbf{P}(B),$$

$$\mathsf{E}\mathbf{I}_{x_i}^2(B) = \mathbf{P}(B), \quad \mathsf{V}\mathbf{I}_{x_i}(B) = \mathbf{P}(B)(1 - \mathbf{P}(B)).$$

Le théorème limite central entraîne immédiatement la proposition suivante:

THÉORÈME 3. *La distribution* $\mathbf{P}_n^*(B)$ *se représente sous la forme*

$$\mathbf{P}_n^*(B) = \mathbf{P}(B) + \frac{\zeta_n(B)}{\sqrt{n}}, \tag{8}$$

où la distribution de $\zeta_n(B) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{I}_{x_i}(B) - \mathbf{P}(B))$ *converge vers la distribution normale de paramètres* $(0, \mathbf{P}(B)(1 - \mathbf{P}(B)))$.

L'étude de $\mathbf{P}_n^*(B)$ sera approfondie dans ce sens au § 6. Des théorèmes de convergence presque sûre plus précis sont accessibles au § 4.

§ 3. Caractéristiques empiriques. Deux types de statistiques

1. Exemples de caractéristiques empiriques. Les *caractéristiques empiriques* sont généralement des fonctionnelles mesurables de la distribution empirique ou, en d'autres termes, des fonctions d'échantillon qui sont supposées mesurables. Les plus simples d'entre elles sont les moments empiriques (ou d'échantillonnage). On appelle *moment empirique d'ordre k* la valeur

$$a_k^* = a_k^*(X) = \int x^k dF_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} x_i^k.$$

Le moment centré d'ordre k est égal à

$$a_k^{*\circ} = a_k^{*\circ}(X) = \int (x - a_1^*)^k dF_n^*(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a_1^*)^k.$$

Les moments empiriques a_1^* et $a_2^{*\circ}$ sont désignés par les symboles spéciaux $\bar{x}$ et S^2 :

$$\bar{x} = a_1^* = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S^2 = a_2^{*\circ} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Dans les problèmes de statistique on se sert de toute sorte de caractéristiques empiriques. Ainsi, la *médiane empirique* ζ^* est la valeur moyenne de la série variationnelle, c'est-à-dire la valeur $\zeta^* = x_{(m)}$ si $n = 2m - 1$ (n est impair), et $\zeta^* = (x_{(m)} + x_{(m+1)})/2$ si $n = 2m$ (n est pair). On rappelle que la médiane ζ d'une distribution continue $\mathbf{P}$ se définit comme la solution de l'équation $F(\zeta) = 1/2$.

Une notion plus générale est celle de quantile ζ_p d'ordre p. On appelle *quantile d'ordre p* le nombre ζ_p tel que $F(\zeta_p) = p$. La médiane est donc le quantile d'ordre 1/2. Si F présente une discontinuité (une composante discrète), cette définition n'a plus de sens. Aussi nous servirons-nous dans le cas général de la définition suivante :

On appelle *quantile ζ_p d'ordre p* de la distribution $\mathbf{P}$ le nombre

$$\zeta_p = \sup \{x : F(x) \leqslant p\}.$$

Traité comme une fonction de p, la quantité ζ_p n'est autre que la fonction $F^{-1}(p)$ inverse de $F(x)$.

Contrairement à la précédente, cette définition de ζ_p (ou de $F^{-1}(p)$) a un sens pour toute $F(x)$.

Il est évident qu'on peut envisager aussi un *quantile empirique ζ_p^* d'ordre p* égal par définition à la valeur $x_{(l)}$, où $l = [np] + 1$ et $x_{(k)}$ sont les termes de l'échantillon ordonné associé à X, $k = 1, \ldots, n$. Pour $p = 1/2$, nous conservons la définition de $\zeta^* = \zeta_{1/2}^*$ donnée ci-dessus (ces deux définitions ne sont confondues que pour les n impairs).

2. Deux types de statistiques. Soit donnée une fonction mesurable S de n arguments. La caractéristique empirique $S(X) = S(x_1, \ldots, x_n)$ est appelée aussi *statistique*. De ce qui précède il est clair que toute statistique est une variable aléatoire, dont la distribution est entièrement définie par la distribution $\mathbf{P}(B) = \mathbf{P}(x_1 \in B)$ (on rappelle que $S(X)$ peut être traitée comme une variable aléatoire définie sur $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}}, \mathbf{P})$, où $\mathbf{P}$ est le n-uple produit direct de distributions de x_1 à une dimension).

Nous distinguerons ici deux classes de statistiques que nous rencontrerons fréquemment dans la suite. Nous les construirons à l'aide des deux

types suivants de fonctionnelles $G(F)$ de la fonction de répartition F :

I. Les fonctionnelles

$$G(F) = h\left(\int g(x)\, dF(x)\right),$$

où g est une fonction borélienne donnée, h une fonction continue au point $a = \int g(x)\, dF_0(x)$, où F_0 est telle que $X \in F_0$.

II. Les fonctionnelles $G(F)$ qui sont continues au « point » F_0 pour la métrique uniforme : $G(F^{(n)}) \to G(F_0)$ si $\sup_x |F^{(n)}(x) - F_0(x)| \to 0$ et les supports *) des distributions $F^{(n)}$ sont contenus dans celui de F_0. Ici F_0 est encore une fonction pour laquelle $X \in F_0$.

Nous définirons les classes de statistiques correspondantes à l'aide de l'égalité

$$S(X) = G(P_n^*),$$

où F_n^* est une fonction de répartition empirique. Nous obtenons alors

I. La classe des *statistiques de type* I

$$S(X) = h\left(\int g(x)\, dF_n^*(x)\right) = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right).$$

Il est évident que tous les moments empiriques sont des statistiques additives $\frac{1}{n}\sum_{i=1}^{n} g(x_i)$ de type I.

II. La classe des *statistiques de type* II ou *statistiques continues au point* F_0.

Il est clair que par exemple la médiane empirique sera une statistique continue au point F si la médiane ζ existe, $F(\zeta) = 1/2$, et F est continue et strictement croissante au point ζ.

Les fonctionnelles n'appartiennent pas nécessairement à l'une ou à l'autre de ces classes. La fonctionnelle $G(F)$ peut n'appartenir à aucune de ces classes ou bien leur appartenir simultanément. Si par exemple G est une fonctionnelle de type I, le support de F est contenu dans l'intervalle $[a, b]$ ($F(a) = 0$, $F(b) = 1$) et la fonction g est à variation bornée sur $[a, b]$, alors G est aussi une fonctionnelle de type II, puisque dans ce cas

$$\int g(x)\, dF(x) = g(b) - \int_a^b F(x)\, dg(x)$$

*) Le support N_F d'une distribution **P** de fonction de répartition F est l'ensemble pour lequel $\mathbf{P}(N_F) = 1$.

est continue par rapport à F pour la métrique uniforme. Ceci exprime que les statistiques $\bar{x}$ et S^2 de type I sont aussi de type II si $X \in \mathbf{P}$ et $\mathbf{P}$ est concentrée sur un intervalle fini.

Les théorèmes 2.1 et 2.2 peuvent être complétés par la proposition suivante relative à la convergence presque sûre des caractéristiques empiriques.

THÉORÈME 1. *Supposons comme précédemment que* $X_n = [X_\infty]_n \in F$. *Si* $S(X) = G(F_n^*)$ *est une statistique de type* I *ou* II, *alors*

$$G(F_n^*) \underset{\text{p.s}}{\to} G(F) \text{ pour } n \to \infty.$$

On admet bien sûr que la valeur $G(F)$ existe.

Donc, les échantillons de grande taille permettent d'estimer non seulement la distribution $\mathbf{P}$, mais aussi les fonctionnelles de cette distribution, du moins celles qui appartiennent à l'une des classes citées dans le théorème.

DÉMONSTRATION. Elle coule de source pour les deux classes de statistiques. Supposons par exemple que $G(F) = h\left(\int g(x)\, dF(x)\right)$. Alors

$$S = S(X) = \int g(x)\, dF_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

est une somme de variables aléatoires indépendantes d'espérance mathématique

$$\mathrm{E}g(x_1) = \int g(x)\, dF(x).$$

Donc, la loi forte des grands nombres nous donne $S \underset{\text{p.s.}}{\to} \mathrm{E}g(x_1)$. Supposons maintenant que $A = \{X_\infty : S(X) \to \mathrm{E}g(x_1)\}$. Alors $\mathbf{P}(A) = 1$ et, si $X_\infty \in A$, il vient $S(X) \to \mathrm{E}g(x_1)$, $h(S(X)) \to h(\mathrm{E}g(x_1))$. En d'autres termes, on a sur l'ensemble A

$$G(F_n^*) \to G(F).$$

Le théorème relatif aux fonctionnelles de type II résulte directement du théorème de Glivenko-Cantelli. ◄

Le théorème 1 nous dit que les moments empiriques centrés et non centrés convergent presque sûrement vers les moments respectifs de $\mathbf{P}$ lorsque $n \to \infty$:

$$a_k^* = a_k^*(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^k \underset{\text{p.s.}}{\to} \mathrm{E}x_1^k,$$

$$a_k^{*\circ} = a_k^{*\circ}(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k \underset{\text{p.s.}}{\to} \mathrm{E}(x_1 - \mathrm{E}x_1)^k.$$

En particulier,

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - \bar{x}^2 \to Vx_1.$$

Nous avons ainsi établi un fait important qui est significatif pour nous : la distribution empirique et une vaste classe de fonctionnelles de cette distribution convergent vers les valeurs « théoriques » correspondantes lorsque $n \to \infty$.

Des théorèmes plus précis sur les distributions des caractéristiques empiriques seront développés aux §§ 7, 8.

§ 4. Echantillons multidimensionnels

1. Distributions empiriques. Les distributions et les caractéristiques empiriques se construisent comme en dimension un lorsque la variable aléatoire ξ, et donc ses valeurs empiriques $x_1, \ldots, x_n$, sont des vecteurs de dimension $m > 1$: $x_k = (x_{k,1}, \ldots, x_{k,m})$. Dans ce cas, $\mathbf{P}(B) = \mathbf{P}(\xi \in B)$ est une distribution sur $\mathscr{X} = R^m$ et $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}}; \mathbf{P})$, où $\mathbf{P}$ est le n-uple produit direct de distributions $\mathbf{P}$ sur $(R^m, \mathfrak{B}_{\mathscr{X}} = \mathfrak{B}^m_R)$, l'espace échantillon. La notation $X \in \mathbf{P}$ reste en vigueur.

La distribution empirique $\mathbf{P}_n^*$ se construit au vu de l'échantillon X comme une distribution discrète de poids $1/n$ en $x_1, \ldots, x_n$, de sorte que

$$\mathbf{P}_n^*(B) = \frac{\nu(B)}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{x_i}(B),$$

où $\nu(B)$ est le nombre de points tombant dans l'ensemble B, $\mathbf{I}_{x_i}$ une distribution concentrée au point x_i.

Le théorème 2.1 de convergence presque sûre de $\mathbf{P}_n^*(B)$ vers $\mathbf{P}(B)$ est manifestement valable.

La généralisation du théorème de Glivenko-Cantelli au cas multidimensionnel est liée à l'émergence de nouveaux problèmes. L'un d'eux est la généralisation de l'intervalle à un rectangle, un ensemble convexe, etc.

La plus simple généralisation du théorème est la suivante.

Soit $y = (y_1, \ldots, y_m)$ un point de R^m et soit B_t un angle de sommet au point $t = (t_1, \ldots, t_m)$:

$$B_t = \{y \in R^m : y_k < t_k,\ k = 1, \ldots, m\}.$$

La fonction

$$F_n^*(t) = \mathbf{P}_n^*(B_t)$$

s'appelle *fonction de répartition empirique*.

THÉORÈME 1. *Soit* $X_n = [X_\infty]_n$, $X_\infty \in F$. *Alors*

$$\sup_t |F_n^*(t) - F(t)| \underset{\text{p.s.}}{\to} 0, \ n \to \infty.$$

2*. Variantes plus générales du théorème de Glivenko-Cantelli. Loi du logarithme itéré. Le théorème de Glivenko-Cantelli admet la généralisation suivante. Soit $\mathfrak{C}$ la classe des ensembles convexes de R^m.

THÉORÈME 2. *Soit* $X_n = [X_\infty]_n$, $X_\infty \in \mathbf{P}$, *et supposons que la distribution* $\mathbf{P}$ *est absolument continue par rapport à la mesure de Lebesgue sur* R^m. *Alors*

$$\sup_{B \in \mathfrak{C}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{p.s.}}{\to} 0. \tag{1}$$

Les autres généralisations éventuelles du théorème 1 peuvent être acquises à l'aide des propositions de l'Annexe I.

REMARQUE 1. La condition de continuité absolue de la distribution $\mathbf{P}$ par rapport à la mesure de Lebesgue est essentielle dans le théorème 2. Ceci est illustré par l'exemple suivant. Supposons que $\mathbf{P}$ est une distribution uniforme sur le cercle unité de R^2. Inscrivons dans ce cercle un polygone fermé B_X de sommets $x_1, \ldots, x_n$. L'ensemble obtenu est convexe. Mais $\mathbf{P}(B_X) = 0$, $\mathbf{P}_n^*(B_X) = 1$, et par suite la relation (1), où $\mathfrak{C}$ est la classe des ensembles convexes, est mise en défaut.

Les théorèmes de type Glivenko-Cantelli peuvent être considérablement affinés, du moins pour les classes élémentaires d'ensembles. Par exemple, pour les fonctions de répartition empiriques $F_n^*(t)$ (cf. théorème 1), on peut exhiber une suite non aléatoire $b_n \to 0$ pour $n \to \infty$, telle que

$$\limsup_{n\to\infty} b_n^{-1} \sup_t |F_n^*(t) - F(t)| = 1$$

presque sûrement (pour presque tous les « points » X_∞). Il s'avère que b_n est du même ordre de petitesse que $\sqrt{\dfrac{\ln \ln n}{n}}$.

THÉORÈME 3 (loi du logarithme itéré). *Si F(t) est continue, on a*

$$\mathbf{P}\left(\limsup_{n\to\infty} \sqrt{\frac{2n}{\ln \ln n}} \sup_t |F_n^*(t) - F(t)| = 1\right) = 1.$$

Le théorème 3 est étroitement lié à l'approximation normale (2.8) de $F_n^*(t)$, qui est valable aussi pour le cas multidimensionnel.

La démonstration des théorèmes 1 et 2 sera donnée dans l'Annexe I. Celle du théorème 3 figure dans [45].

3. Caractéristiques empiriques. En dimension un comme en dimension $m > 1$ ce sont des fonctions d'échantillon mesurables. Les plus élémentaires

d'entre elles sont les moments empiriques. Par exemple, les moments empiriques d'ordre un sont égaux à

$$a_{1,j}^{\bullet} = a_{1,j}^{\bullet}(X) = \frac{1}{n}\sum_{k=1}^{n} x_{k,j}, \quad j = 1, \ldots, m.$$

Les moments d'ordre deux centrés et non centrés sont

$$a_{2,ij}^{\bullet} = a_{2,ij}^{\bullet}(X) = \frac{1}{n}\sum_{k=1}^{n} x_{k,i}x_{k,j}, \quad i, j = 1, \ldots, m,$$

$$a_{2,ij}^{\bullet\circ} \equiv S_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{k,i} - a_{1,i}^{\bullet})(x_{k,j} - a_{1,j}^{\bullet}),$$

etc. Comme en dimension un, on s'assure sans peine, en appliquant la loi forte des grands nombres, que ces caractéristiques convergent presque sûrement vers les moments « théoriques » respectifs. En particulier, $S_{ij} \underset{\text{p.s.}}{\rightarrow} \mathrm{E}(x_{1,i} - \mathrm{E}x_{1,i})(x_{1,j} - \mathrm{E}x_{1,j})$. On s'assure aisément (pour plus de détails voir le paragraphe suivant) que les coefficients de corrélation empiriques

$$r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}} \underset{\text{p.s.}}{\rightarrow} \varrho(x_{1,i}, x_{1,j}) = \frac{\mathrm{E}(x_{1,i} - \mathrm{E}x_{1,i})(x_{1,j} - \mathrm{E}x_{1,j})}{\sqrt{\mathrm{V}x_{1,i}\mathrm{V}x_{1,j}}}.$$

possèdent aussi cette propriété.

Les théorèmes de continuité qui vont suivre nous permettront d'établir des propositions plus précises sur la distribution des caractéristiques empiriques.

§ 5. Théorèmes de continuité

Pour la suite de l'exposé nous aurons besoin de propositions auxiliaires qu'on pourrait appeler *théorèmes de continuité*. Pour la commodité nous les regrouperons en un seul paragraphe. Nous avons déjà utilisé un de ces théorèmes, savoir le théorème 3.1. Le premier théorème de continuité en est très proche.

THÉORÈME 1 (premier théorème de continuité). *Soit* $X = [X_\infty]_n \in \mathbf{P}$. *Si* $S_n = S_n(X)$ *est une suite de statistiques scalaires ou vectorielles telles que* $S_n \underset{\text{p.s.}}{\rightarrow} S_0$ *et* $H(s)$ *est une fonction continue presque partout par rapport à la distribution de la variable aléatoire* S_0 *(autrement dit* $H(s)$ *est continue en chaque point de l'ensemble* B, $\mathbf{P}(S_0 \in B) = 1$*), alors* $H(S_n(X)) \underset{\text{p.s.}}{\rightarrow} H(S_0)$.

Si S_n converge en probabilité vers S_0 (on notera ceci $S_n \underset{P}{\to} S_0$), alors pour les mêmes conditions $H(S_n) \underset{P}{\to} H(S_0)$.

DÉMONSTRATION. Elle est presque évidente. Puisque les événements $A = \{X_\infty : S_n(X_\infty) \to S_0(X_\infty)\}$ et $C = \{X_\infty : S_0(X_\infty) \in B\}$ sont presque sûrs, il en est de même, en vertu de l'égalité $\mathbf{P}(A \cap C) = \mathbf{P}(A) + \mathbf{P}(C) - \mathbf{P}(A) \cup \cup C)$, de l'événement $A \cap C$ (sur lequel $H(S_n(X_\infty)) \to H(S_0(X_\infty))$).

Pour alléger la démonstration de la convergence en probabilité, nous admettrons accessoirement que $S_0 =$ const (nous n'aurons besoin que de ce cas). Pour $\epsilon > 0$, il existe un $\delta > 0$ tel que l'événement $A_n = \{X_\infty : |S_n - - S_0| < \delta\}$ entraîne $|H(S_n) - H(S_0)| < \epsilon$ et de plus $\mathbf{P}(A_n) > 1 - \epsilon$ pour tous les n assez grands. Donc, pour de tels n on a $1 - \epsilon < \mathbf{P}(A_n) \leqslant \leqslant \mathbf{P}(|H(S_n) - H(S_0)| < \epsilon)$. ◀

Avant de formuler les théorèmes suivants introduisons quelques notations.

Soit donnée une suite de vecteurs aléatoires $\eta_n = (\eta_n^{(1)}, \ldots, \eta_n^{(s)})$ (pas nécessairement sur le même espace probabilisé). Si les distributions de η_n convergent faiblement pour $n \to \infty$ vers la distribution d'une variable aléatoire η, on notera ce fait par le symbole

$$\eta_n \Rightarrow \eta. \tag{1}$$

Nous utiliserons ce symbole également pour les distributions, de sorte que la relation (1) équivaut à

$$\mathbf{Q}_n \Rightarrow \mathbf{Q},$$

où $\mathbf{Q}_n$ et $\mathbf{Q}$ sont les distributions respectives de η_n et de η. Cette convention est commode et pas ambiguë.

Il est clair que de $\eta_n \underset{P}{\to} \eta$ ou $\eta_n \underset{\text{p.s.}}{\to} \eta$ il s'ensuit que $\eta_n \Rightarrow \eta$ (comparer avec [11]).

Si donc il est question d'une relation (mettant en jeu la convergence faible) entre objets de même nature (entre variables aléatoires ou entre distributions), on se servira du symbole $\Rightarrow$. Il serait commode de disposer aussi d'un symbole pour exprimer que « *les distributions de η_n convergent faiblement vers $\mathbf{Q}$ lorsque $n \to \infty$* ». On notera ceci par

$$\eta_n \Subset \mathbf{Q}, \tag{2}$$

de sorte que le symbole $\Subset$ exprime le même fait que le symbole $\Rightarrow$ mais pour des objets de nature différente (à gauche on a des variables aléatoires, à droite, une distribution).

Soient η_n et η des vecteurs aléatoires de R^s.

THÉORÈME 2 (deuxième théorème de continuité). *Si $\eta_n \Rightarrow \eta$ et $H(t)$, $t \in \in R^s$, est une fonction continue de R^s dans R^k, alors $H(\eta_n) \Rightarrow H(\eta)$.*

Signalons que ce théorème est en fait valable dans une forme plus générale *) : *Si* $\eta_n \Rightarrow \eta$ *et* $H(t)$ *est continue dans un ensemble* $A \in \mathfrak{B}^s$, $\mathbf{P}(\eta \in A) = 1$, *alors* $H(\eta_n) \Rightarrow H(\eta)$.

DÉMONSTRATION du théorème 2. Supposons que $\mathbf{Q}_n$ et $\mathbf{Q}$ sont les distributions respectives de η_n et de η. La convergence faible de $\mathbf{Q}_n$ vers $\mathbf{Q}$ exprime par définition que pour toute fonction continue et bornée $f: R^s \to R$, on a

$$\int f(y)\mathbf{Q}_n(dy) \to \int f(y)\mathbf{Q}(dy),$$

ou, ce qui est équivalent,

$$\mathbf{E}f(\eta_n) \to \mathbf{E}f(\eta). \tag{3}$$

Nous devons obtenir une relation identique pour les distributions de $H(\eta_n)$ et de $H(\eta)$. Autrement dit, nous devons établir que pour toute fonction bornée continue $g: R^k \to R$, on a $\mathbf{E}g(H(\eta_n)) \to \mathbf{E}g(H(\eta))$. Or ceci résulte directement de (3), puisque la composée $\tilde{g} \equiv g \circ H: R^s \to R$ est continue et bornée. ◄

THÉORÈME 3 (troisième théorème de continuité). *Soit* $\eta_n \Rightarrow \eta \in R$ *et soit* $H(t)$, $t \in R$, *une fonction dérivable au point* a. *Si* b_n *est une suite numérique convergeant vers* 0, *alors*

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta H'(a). \tag{4}$$

DÉMONSTRATION. Considérons la fonction

$$h(x) = \begin{cases} (H(a + x) - H(a))/x, & x \neq 0, \\ H'(a), & x = 0, \end{cases}$$

qui est continue au point $x = 0$. Puisque $b_n\eta_n \Rightarrow 0$, le premier théorème de continuité nous donne $h(b_n\eta_n) \Rightarrow h(0) = H'(a)$. Le deuxième théorème de continuité entraîne

$$(H(a + b_n\eta_n) - H(a))/b_n = h(b_n\eta_n)\eta_n \Rightarrow H'(a)\eta. ◄$$

Citons maintenant deux généralisations du théorème 3.

THÉORÈME 3A. *Soit* $\eta_n \equiv (\eta_n^{(1)}, \ldots, \eta_n^{(s)}) \Rightarrow \eta \equiv (\eta^{(1)}, \ldots, \eta^{(s)})$ *et soit* $H(t)$ *une fonction scalaire du vecteur* $t = (t_1, \ldots, t_s)$, *dont la dérivée* $H'(t) \equiv \left(\frac{\partial H}{\partial t_1}, \ldots, \frac{\partial H}{\partial t_s}\right)$ *existe en* a. *Si* $b_n \to 0$, *on a alors*

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta(H'(a))^T = \sum_{j=1}^{s} \frac{\partial H(a)}{\partial t_j}\eta^{(j)}. \tag{5}$$

L'indice T représente la transposition.

---

*) Pour plus de détails voir [5].

Si $\eta(H'(a))^T = 0$ *presque sûrement (par exemple* $H'(a) = 0$), *et si la matrice* $H''(t)$ *des dérivées* $\dfrac{\partial^2 H(t)}{\partial t_i \partial t_j}$ *existe en* a, *alors*

$$(H(a + b_n\eta_n) - H(a))/b_n^2 \Rightarrow \frac{1}{2}\,\eta H''(a)\eta^T = \frac{1}{2}\sum_{i,j=1}^{s} \frac{\partial^2 H(a)}{\partial t_i \partial t_j}\,\eta^{(i)}\eta^{(j)}. \quad (6)$$

Supposons maintenant que $H(t)$ est une fonction vectorielle. Il est alors évident que la distribution limite de chaque composante H_j sera décrite par le théorème 3A et que la distribution conjointe sera justiciable du

THÉORÈME 3B. *Soit* $\eta_n \Rightarrow \eta \in R^s$ *et soit* $H(t) \in R^k$ *une fonction vectorielle dont les dérivées* H_j', $j = 1, \ldots, k$, *vérifient les conditions du théorème* 3A. *Alors*

$$(H(a + b_n\eta_n) - H(a))/b_n \Rightarrow \eta(H'(a))^T.$$

Si $\eta(H'(a))^T = 0$ *presque sûrement et que les matrices* H_j'', $j = 1, \ldots, k$, *existent au point* a, *alors*

$$(H(a + b_n\eta_n) - H(a))/b_n^2 \Rightarrow \frac{1}{2}\,(\eta H_1''(a)\eta^T, \ldots, \eta H_k''(a)\eta^T).$$

Les démonstrations de ces théorèmes sont pratiquement les mêmes que celle du théorème 3, c'est pourquoi nous les proposons au lecteur à titre d'exercice. Nous proposons par ailleurs au lecteur de s'assurer que dans (4), (5) et (6) il est possible de remplacer le symbole $\Rightarrow$ par $\underset{p.s.}{\to}$ ou $\underset{P}{\to}$ si respectivement $\eta_n \underset{p.s.}{\to} \eta$ ou $\eta_n \underset{P}{\to} \eta$.

Les théorèmes 1, 2 et 3 se résument de la manière suivante. Supposons que $\rightsquigarrow$ désigne l'un des symboles $\underset{p.s.}{\to}$, $\underset{P}{\to}$, $\Rightarrow$. Si H est continue, alors $\eta_n \rightsquigarrow \eta$ entraîne $H(\eta_n) \rightsquigarrow H(\eta)$.

Si H est dérivable au point a, et $\eta_n \rightsquigarrow \eta$, alors pour $b_n \to 0$, on a

$$(H(a + b_n\eta_n) - H(a))/b_n \rightsquigarrow H'(a)\eta. \quad (7)$$

REMARQUE 1. Il est immédiat de voir que si a dépend de n de telle sorte que $a \equiv a_n = a_0 + o(1)$ et si les dérivées figurant dans les théorèmes 3, 3A et 3B sont continues, la relation (7) reste en vigueur sous la forme

$$(H(a_n + b_n\eta_n) - H(a_n))/b_n \rightsquigarrow H'(a_0)\eta. \quad (8)$$

Pour le prouver, il suffit de remarquer que le premier membre de (8) se représente sous la forme $H'(\alpha_n)\eta_n$, où $\alpha_n = \theta a_n + (1 - \theta)(a_n + b_n\eta_n) \rightsquigarrow a_0$, $|\theta| \leqslant 1$, et d'appliquer le deuxième théorème de continuité.

Cette remarque est également valable pour les analogues multidimensionnels de cette proposition (théorèmes 3A et 3B).

Les théorèmes formulés concernaient la convergence presque sûre et la convergence des distributions. Le quatrième théorème porte sur la convergence d'intégrales.

THÉORÈME 4 (théorème de continuité des moments). *Soit* $\{\eta_n\}$ *une suite de variables aléatoires numériques. Supposons que* $\eta_n \Rightarrow \eta$ *lorsque* $n \to \infty$. *Dans ces conditions, si l'une au moins des conditions suivantes*

$$1)\ \lim_{n\to\infty}\sup \int_N^\infty \mathbf{P}(|\eta_n|>x)dx \to 0 \textit{ pour } N\to\infty,$$

$$2)\ \mathbf{P}(|\eta_n|>x) \leqslant \varphi(x),\ \int_0^\infty \varphi(x)dx < \infty,$$

$$3)\ \mathsf{E}|\eta_n|^{1+\alpha} < c < \infty \textit{ pour un } \alpha > 0,$$

est réalisée, alors $\lim\limits_{n\to\infty} \mathsf{E}\eta_n = \mathsf{E}\eta$.

Signalons que la condition 1 exprime que $\int_N^\infty \mathbf{P}(|\eta_n|>x)dx$ tend vers 0 uniformément en n lorsque $N\to\infty$.

DÉMONSTRATION. L'inégalité généralisée de Tchébychev

$$\mathbf{P}(|\eta_n|>x) \leqslant \frac{\mathsf{E}|\eta_n|^{1+\alpha}}{x^{1+\alpha}}$$

nous dit que la condition 3 entraîne la condition 2. A son tour la condition 2 entraîne la condition 1.

Supposons que la condition 1 est réalisée. Pour alléger les raisonnements on admettra d'abord que $\eta_n \geqslant 0$. Une intégration par parties nous donne alors

$$\mathsf{E}\eta_n = -\int_0^\infty x\, d\mathbf{P}(\eta_n \geqslant x) = \int_0^\infty \mathbf{P}(\eta_n \geqslant x)dx.$$

Cette représentation, la convergence $\mathbf{P}(\eta_n \geqslant x) \to \mathbf{P}(\eta \geqslant x)$ pour presque tous les x et la convergence uniforme en n de l'intégrale $\int_0^\infty \mathbf{P}(\eta_n \geqslant x)dx$ entraînent la légitimité du passage à la limite sous le signe d'intégration, soit

$$\lim_{n\to\infty} \mathsf{E}\eta_n = \lim_{n\to\infty}\int_0^\infty \mathbf{P}(\eta_n \geqslant x)dx = \int_0^\infty \mathbf{P}(\eta \geqslant x)dx = \mathsf{E}\eta.$$

Dans le cas général, il faut se servir de la représentation $\eta_n = \eta_n^+ - \eta_n^-$, où $\eta_n^+ = \max(\eta_n, 0)$, $\eta_n^- = \max(-\eta_n, 0)$. ◄

Signalons que la condition 1 peut être traitée aussi comme une condition d'intégrabilité uniforme de η_n, qui entraîne immédiatement la convergence annoncée. $\mathsf{E}\eta_n \to \mathsf{E}\eta$ (cf. par exemple [11], [52]).

§ 6*. Fonction de répartition empirique en tant que processus aléatoire. Convergence vers le pont brownien

Dans ce paragraphe on admettra connue la notion de *processus aléatoire* et en particulier les définitions et les propriétés élémentaires des *processus wienérien* et *poissonnien.*

1. Distribution du processus $nF_n^*(t)$. On limitera notre étude au cas de la dimension un, c'est-à-dire au cas où $\mathscr{X} = R$. Supposons comme précédemment que $F_n^*(t) = \mathbf{P}_n^*]-\infty, t[$ est une fonction de répartition empirique associée à un échantillon $X = X_n \in \mathbf{P}$.

La fonction $F_n^*(t)$ est une fonction de deux variables : t et X, ou, ce qui revient au même, une fonction aléatoire de t ou un *processus aléatoire.*

Trouvons les *distributions finidimensionnelles* de ce processus. Soient $t_1 < t_2 < \ldots < t_m$ m points arbitraires de la droite numérique. Posons $t_0 = -\infty$, $t_{m+1} = \infty$ et désignons par

$$\Delta_j g = g(t_{j+1}) - g(t_j)$$

les accroissements de la fonction $g(t)$ sur les semi-intervalles $\Delta_j = [t_j, t_{j+1}[$, $j = 0, 1, \ldots, m$. Considérons l'accroissement $\Delta_j \pi_n$ du processus

$$\pi_n(t) = nF_n^*(t).$$

Il est évident que c'est le nombre d'éléments de l'échantillon qui tombent dans l'intervalle Δ_j. La probabilité d'entrée d'un élément de l'échantillon (disons x_1) dans Δ_j est égale à $p_j = \mathbf{P}(\Delta_j)$. Vu que les entrées dans Δ_j, $j = 0, 1, \ldots, m$, sont des événements incompatibles, le vecteur $(\Delta_0 \pi_n, \ldots, \Delta_m \pi_n)$ admet visiblement une distribution polynomiale (cf. [11]) avec les probabilités $p_0, \ldots, p_m$, $\sum_{j=0}^{m} p_j = 1$. On sait que

$$\mathbf{P}(\Delta_0 \pi_n = k_0, \ldots, \Delta_m \pi_n = k_m) = \frac{n!}{k_0! \ldots k_m!} p_0^{k_0} \ldots p_m^{k_m}, \tag{1}$$

où

$$\sum_{j=0}^{m} k_j = n.$$

Supposons maintenant que $\eta(u)$, $u \in [0, 1]$, est un *processus poissonnien* continu à gauche (cf. [11]) *de paramètre* λ, $\eta(0) = 0$. Les accroissements de ce processus sont indépendants

$$\mathbf{P}(\eta(u) = k) = e^{-\lambda u} \frac{(\lambda u)^k}{k!}.$$

Si la fonction de répartition $F(t) = \mathbf{P}]-\infty, t[$ est continue, nous pouvons faire un changement de temps continu en posant $u = F(t)$, $t \in]-\infty, \infty[$, et définir ainsi le processus $\pi(t) = \eta(F(t))$ sur l'axe temporel

tout entier. Considérons les accroissements

$$\Delta_j \pi = \pi(t_{j+1}) - \pi(t_j) = \eta(F(t_{j+1})) - \eta(F(t_j))$$

de ce processus sur les intervalles Δ_j. Alors

$$\mathbf{P}(\Delta_0 \pi = k_0, \ldots, \Delta_m \pi = k_m) = \prod_{j=0}^{m} e^{-\lambda p_j} \frac{(\lambda p_j)^{k_j}}{k_j!} = e^{-\lambda} \lambda^n \prod_{j=0}^{m} \frac{p_j^{k_j}}{k_j!},$$

quant à la probabilité conditionnelle de cet événement (sachant que $\pi(\infty) = \sum_{j=0}^{m} \Delta_j \pi = n$), elle sera égale à

$$\mathbf{P}\left(\Delta_0 \pi = k_0, \ldots, \Delta_m \pi = k_m \middle| \sum_{j=0}^{m} \Delta_j \pi = n\right) =$$

$$= \frac{\mathbf{P}(\Delta_0 \pi = k_0, \ldots, \Delta_m \pi = k_m)}{\mathbf{P}(\pi(\infty) = n)} = \mathbf{P}(\Delta_0 \pi = k_0, \ldots, \Delta_m \pi = k_m) \frac{e^{\lambda} n!}{\lambda^n} =$$

$$= n! \prod_{j=0}^{m} \frac{p_j^{k_j}}{k_j!}. \qquad (2)$$

Pour tout $\lambda > 0$, nous avons obtenu la même expression qu'au second membre de (1). Nous avons ainsi prouvé le

THÉORÈME 1. *Si $F(t)$ est une fonction continue, la distribution du processus $nF_n^*(t)$ est confondue avec la distribution conditionnelle du processus $\pi(t) = \eta(F(t))$ sachant que $\pi(\infty) = n$ ($\eta(1) = n$).*

Le théorème exprime que les écarts $n(F_n^*(t) - F(t))$ sont distribués comme $\eta(F(t)) - nF(t)$ sachant que $\eta(1) = n$, et le problème se ramène, au changement $u = F(t)$ près, à l'étude des écarts $\eta(u) - nu$ d'un processus poissonnien conditionnel (sachant que $\eta(1) = n$) sur l'intervalle $[0, 1]$ ou, ce qui est équivalent, à l'étude des écarts $n(F_n^*(t) - t)$, où $F_n^*(t)$ correspond à une distribution uniforme sur l'intervalle $[0, 1]$.

Le processus $nF_n^*(t)$ admet une autre représentation utile pour les applications. Soient $\zeta_1, \zeta_2, \ldots$, les points de discontinuité d'un processus poissonnien $\eta(t)$: $\eta(\zeta_k + 0) = k$. On sait que les différences $\xi_k = \zeta_k - \zeta_{k-1}$ ($\zeta_0 = 0$), $k = 1, 2, \ldots$, sont indépendantes et exponentiellement distribuées :

$$\mathbf{P}(\xi_k > x) = e^{-\lambda x},$$

ζ_k suit une loi gamma de densité (cf. aussi § 2.2)

$$\gamma_{\lambda,k}(x) = \frac{\lambda^k}{\Gamma(k)} e^{-\lambda x} x^{k-1}.$$

Pour alléger les énoncés, on supposera que $F(t) = t$, $t \in [0, 1]$, $t_0 = 0$, $t_{m+1} = 1$, de sorte que $\eta(t) = \pi(t)$.

THÉORÈME 2. *La distribution du processus* $nF_n^*(t)$ *est confondue pour tout* $v > 0$ *avec la distribution conditionnelle du processus* $\pi(tv)$, $t \in]0, 1[$, *sachant que* $\zeta_{n+1} = v$.

Autrement dit, le théorème 1 reste en vigueur si l'on remplace la condition $\pi(1) = n$ par la condition bien plus restrictive $\pi(1) = n$, $\pi(1 + 0) = n + 1$ (on admet que les trajectoires $\pi(t)$ sont continues à gauche).

Vu que la probabilité de cette nouvelle condition est nulle, il convient d'ajouter (cf. §§ 4, 8 de [11] sur les espérances mathématiques conditionnelles, ainsi que le § 2.9) que par distribution conditionnelle on comprend les probabilités

$$\mathbf{P}(A \mid \zeta_{n+1} = v) = \frac{\mathbf{P}(A;\ \zeta_{n+1} \in dv)}{\mathbf{P}(\zeta_{n+1} \in dv)},$$

où $A = \{\Delta_0\pi(tv) = k_0, \ldots, \Delta_m\pi(tv) = k_m\}$, $\Delta_j\pi(tv) = \pi(t_{j+1}v) - \pi(t_jv)$.

DÉMONSTRATION. L'événement $\{\zeta_{n+1} \in dv\}$ peut être représenté par le produit des deux événements

$$B = \{\pi(v) = n\} \quad \text{et} \quad C = \{\pi(v + dv) - \pi(v) = 1\}.$$

Les événements B et AB ne dépendent pas de C, puisque les événements B et AB d'une part et l'événement C de l'autre font partie des accroissements du processus π sur des intervalles de temps disjoints. Donc

$$\mathbf{P}(A \mid \zeta_{n+1} = v) = \frac{\mathbf{P}(ABC)}{\mathbf{P}(BC)} = \frac{\mathbf{P}(AB)}{\mathbf{P}(B)} = \mathbf{P}(A \mid \pi(v) = n). \qquad (3)$$

Exactement comme dans (2), on s'assure que cette expression ne dépend ni de v ni de λ et est confondue avec (1). ◄

COROLLAIRE 1. *La distribution du processus* $nF_n^*(t)$ *est confondue avec celle de* $\pi(t\zeta_{n+1})$, $t \in [0, 1]$.

Ceci résulte du fait que pour $B = \{\Delta_0\pi(t\zeta_{n+1}) = k_0, \ldots, \Delta_m\pi(t\zeta_{n+1}) = k_m\}$, on a en vertu de (3)

$$\mathbf{P}(B) = \int \mathbf{P}(A \mid \zeta_{n+1} = v)\mathbf{P}(\zeta_{n+1} \in dv) = n! \prod_j \frac{\Delta_j^{k_j}}{k_j!}.$$

Le corollaire 1 entraîne le

COROLLAIRE 2. *La distribution conjointe des éléments de l'échantillon ordonné* $x_{(1)}, \ldots, x_{(n)}$ *associé à un échantillon* X *de distribution uniforme*

est confondue avec la distribution conjointe de

$$\frac{\zeta_1}{\zeta_{n+1}}, \ldots, \frac{\zeta_n}{\zeta_{n+1}},$$

ou, ce qui revient au même, la distribution conjointe des différences $x_{(1)}$, $x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(n-1)}, 1 - x_{(n)}$ *est confondue avec celle de*

$$\frac{\xi_1}{\zeta_{n+1}}, \ldots, \frac{\xi_{n+1}}{\zeta_{n+1}}.$$

Nous allons achever ce n° par le calcul des moments d'ordre deux des accroissements du processus $n(F_n^*(t) - F(t))$. Il nous sera plus commode de traiter le processus

$$w^n(t) = \sqrt{n}(F_n^*(t) - F(t)).$$

Il est évident que $\mathsf{E}\Delta_j w^n = 0$, $\mathsf{E}(\Delta_j w^n)^2 = \Delta_j F(1 - \Delta_j F)$. Pour calculer les moments mixtes, on remarquera que $(i \neq j)$

$$\mathsf{E}(\Delta_i w^n \cdot \Delta_j w^n) = \frac{1}{2} \sum_{k,l=1}^{n} \mathsf{E}(\mathbf{I}_{x_k}(\Delta_i) - \mathbf{P}(\Delta_i))(\mathbf{I}_{x_l}(\Delta_j) - \mathbf{P}(\Delta_j)) =$$

$$= \frac{1}{n} \sum_{k,l=1}^{n} \{\mathbf{E}\mathbf{I}_{x_k}(\Delta_i)\mathbf{I}_{x_l}(\Delta_j) - \mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j)\}.$$

Comme

$$\mathbf{E}\mathbf{I}_{x_k}(\Delta_i)\mathbf{I}_{x_l}(\Delta_j) = \begin{cases} \mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j) & \text{si} \quad k \neq l, \\ 0 \quad \text{si} \quad k = l, \end{cases}$$

il vient

$$\mathsf{E}(\Delta_i w^n \cdot \Delta_j w^n) = -\mathbf{P}(\Delta_i)\mathbf{P}(\Delta_j) = -\Delta_i F \cdot \Delta_j F.$$

Donc, les accroissements du processus w^n sont négativement corrélés.

2. Comportement du processus $w^n(t)$ à la limite. On admettra que $F(t)$ est continue. Du n° 1 il s'ensuit alors qu'on peut se borner à l'étude d'une distribution $F(t) = t$, $t \in [0, 1]$, uniforme sur $[0, 1]$.

Désignons par $w(t)$ un *processus wienérien standard*, c'est-à-dire un processus à accroissements indépendants dont les valeurs suivent une loi normale de paramètres $(0, t)$. Le processus

$$w^\circ(t) = w(t) - tw(1)$$

s'appelle *pont brownien* (car ces deux extrémités sont fixées : $w^\circ(0) = w^\circ(1) = 0$). La distribution de ce processus est confondue avec la distribution conditionnelle du processus $w(t)$ sachant que $w(1) = 0$ (plus exactement, il faut prendre la condition $|w(1)| < \epsilon$ et passer à la limite pour $\epsilon \to 0$).

Il s'avère que les distributions finidimensionnelles des processus

$$w^n(t) = \sqrt{n}(F_n^*(t) - F(t)), \quad t \in [0, 1],$$

convergent pour $n \to \infty$ vers les distributions respectives du pont brownien $w°(t)$.

Ce fait permet d'approcher les processus $w^n(t)$, appelés parfois *processus empiriques*, par le processus $w°(t)$. Plus exactement, on peut concevoir que pour de grands n, on a la relation

$$\sqrt{n}(F_n^*(t) - F(t)) \approx w°(t), \tag{4}$$

qui décrit la distribution des écarts entre $F_n(t)$ et $F(t)$ (on rappelle qu'on a convenu que $F(t) = t$, $t \in [0, 1]$).

Mais nous aurons besoin d'une relation (4) plus forte. Considérons par exemple la statistique $U = \sqrt{n} \sup_t (F_n^*(t) - F(t))$. La relation (4) nous invite tout naturellement à supposer que pour les grands n, la variable aléatoire U est distribuée approximativement comme $\sup_{0 \le t \le 1} w°(t)$. Or ceci ne résulte en aucune façon de notre relation (4), puisque U ne peut être représentée comme une fonction des valeurs de $w^n(t) = \sqrt{n}(F_n^*(t) - F(t))$ en un nombre fini de points. Donc la proposition suivante est bien plus forte.

Désignons par $D(a, b)$ l'espace des fonctions définies sur $[a, b]$, continues à gauche (au point a à droite) et présentant un nombre fini de sauts, et par $C(a, b)$ l'espace de toutes les fonctions continues sur $[a, b]$. Il est évident que les trajectoires $w^n(t)$ appartiennent à $D(0, 1)$. On sait par ailleurs (cf. [11] chap. 13) que la trajectoire $w°(t)$ appartient presque sûrement à $C(0, 1)$. Par souci de simplicité, on peut admettre que toutes les trajectoires $w(t)$, et partant la trajectoire $w°(t)$, sont contenues dans $C(0, 1)$ (cf. [11]). Comme $C(0, 1) \subset D(0, 1)$, il s'ensuit que $(D(0, 1), \sigma_D)$, où σ_D est la tribu des sous-ensembles cylindriques *) de $D(0, 1)$, peut être considéré comme l'espace échantillon **) des processus w^n et $w°$.

THÉORÈME 3 (théorème limite fonctionnel pour les processus empiriques). *Soit f une fonctionnelle définie sur l'espace $D(0, 1)$, telle que*

1) *$f(w^n)$ et $f(w°)$ soient des variables aléatoires (c'est-à-dire que $f(y)$ soit une application mesurable de $(D(0, 1), \sigma_D)$ dans $(R, \mathfrak{B})$)* ;

2) *$f(y)$ soit une fonctionnelle continue aux « points » de l'espace*

---

*) C'est-à-dire des ensembles de la forme $\{y(t_1) \in B_1, \ldots, y(t_m) \in B_m\}$, où $B_1, \ldots, B_m$ sont des boréliens.

**) (D_0, σ) est l'espace échantillon du processus $\xi(t)$ si la distribution de ξ est définie sur lui, si bien que les trajectoires $\xi(t)$ sont contenues dans D_0.

$C(0, 1)$ pour la métrique uniforme, c'est-à-dire que $f(y_n) \to f(y)$ pour $n \to \infty$, pourvu que $y \in C(0, 1)$ et $\varrho(y_n, y) = \sup_{0 \le t \le 1} |y_n(t) - y(t)| \to 0$.

Si ces conditions sont remplies, alors

$$f(w^n) \Rightarrow f(w^\circ).$$

Si la fonctionnelle f est continue pour la métrique uniforme en tout point $y \in D(0, 1)$, la condition 1) est automatiquement réalisée.

Il est évident que la fonctionnelle U envisagée ci-dessus remplit les conditions du théorème, de sorte que pour $n \to \infty$

$$U \Rightarrow \sup_{0 \le t \le 1} w^\circ(t).$$

Puisque la distribution du second nombre de cette relation peut être trouvée sous une forme explicite (cf. par exemple [5], [75]) :

$$\mathbf{P}(\sup_{0 \le t \le 1} w^\circ(t) > z) = e^{-2z^2},$$

on obtient ainsi une expression approchée de la distribution de U.

Dans les paragraphes suivants, le théorème 3 est utilisé pour calculer la distribution limite d'autres statistiques.

La démonstration du théorème 3 est reportée à l'Annexe II.

§ 7. Distribution limite des statistiques du premier type

On rappelle qu'une statistique du premier type est une statistique $S_n(X) = G(F_n^*)$, où la fonctionnelle $G(F) = h\left(\int g(x) dF(x)\right)$. Autrement dit

$$S_n(X) = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right).$$

Nous avons vu (théorème 3.1) que si $X \in F_0$ et si h est continue en $a = \int g(x)\, dF_0(x)$, alors $S_n \underset{\text{p.s.}}{\to} h(a)$.

THÉORÈME 1. *Si $X \in F_0$, h est dérivable au point a et $\int g^2(x)\, dF_0(x) < \infty$, alors*

$$\sqrt{n}(S_n(X) - h(a)) \Rightarrow h'(a)\xi,$$

où $\xi \in \Phi_{0,\sigma^2}$, Φ_{0,σ^2} désigne la distribution normale de paramètres $(0, \sigma^2)$, et $\sigma^2 = \int (g(x) - a)^2 dF_0(x)$.

DÉMONSTRATION. Mettons la statistique $S_n(X)$ sous la forme

$$h\left(a + \frac{1}{\sqrt{n}}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - a)\right]\right),$$

où, en vertu du théorème limite central (cf. [11]),

$$\eta_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - a) \in \Phi_{0,\sigma^2},$$

$$\sigma^2 = \mathbf{E}(g(x_1) - a)^2 = \int(g(x) - a)^2 dF_0(x).$$

Il reste à appliquer le troisième théorème de continuité pour $b_n = 1/\sqrt{n}$. ◄

On aura parfois intérêt à étudier les fonctionnelles du premier type sous la forme $G(F) = h\left(\int g(x)d(F - F_0)\right)$. Il est évident qu'elles sont justiciables de tout ce qui précède, à la seule différence qu'il faut poser $a = 0$.

Citons un analogue du théorème 1 pour le cas où la fonction $g = (g_1, \ldots$ $\ldots, g_s)$ est vectorielle (c'est-à-dire que $G(F) = h\left(\int g_1(x)dF(x), \ldots\right.$ $\left.\ldots, \int g_s(x)dF(x)\right)$).

THÉORÈME 1A. *Supposons que* $S_n(X) = G(F_n^*)$, $h(t)$ *est dérivable au point* $a = \int g(x)dF_0(x)$ *et la matrice des moments d'ordre deux* $\sigma^2 = \|\sigma_{ij}\| =$ $= \mathbf{E}(g(x_1) - a)^T(g(x_1) - a)$ *est finie. Alors*

$$(S_n(X) - h(a))\sqrt{n} \Rightarrow \xi(h'(a))^T = \sum_{j=1}^{s}\frac{\partial h(a)}{\partial t_j} = \xi_j, \qquad (1)$$

où $\xi = (\xi_1, \ldots, \xi_s) \in \Phi_{0,\sigma^2}$.

Si $\xi(h'(a))^T = 0$ *presque sûrement et la matrice des dérivées secondes* $h''(t) = \left\|\frac{\partial^2}{\partial t_i \partial t_j} h(t)\right\|$ *existe au point a, alors*

$$(S_n(X) - h(a))n \Rightarrow \frac{1}{2}\xi h''(a)\xi^T = \frac{1}{2}\sum_{i,j=1}^{s}\frac{\partial^2 h(t)}{\partial t_i \partial t_j}\xi_i\xi_j.$$

Pour prouver le théorème 1A, il faut se servir du théorème de continuité 5.3A et du théorème limite central multidimensionnel en vertu duquel $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(x_i) - a) \Rightarrow \xi$ (cf. Annexe V).

Le théorème de la distribution limite de $S_n(X)$ s'énonce exactement dans les mêmes termes lorsque la fonction h, donc la statistique $S_n(X)$, sont des vecteurs. Le lecteur aura la partie belle de produire l'énoncé et la démonstration à l'aide du théorème 5.3B.

EXEMPLE 1. Supposons que $X \in \mathbf{P}_0$ et que $\mathbf{P}_0$ est telle que $\mathbf{E}x_1 = \alpha > 0$, $\mathbf{V}x_1 = d^2 < \infty$. On demande la distribution limite de la statistique $S =$ $= 1/\bar{x} \left(\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\right)$. Les conditions du théorème 1 sont visiblement remplies pour $h(t) = 1/t$, $g(x) = x$, et de plus $a = \alpha$, $\sigma^2 = d^2$, $h(a) = 1/\alpha$, $h'(a) = -1/\alpha^2$. D'après le théorème 1,

$$(S - 1/\alpha)\sqrt{n} \Rightarrow -\xi/\alpha^2, \quad \xi \in \Phi_{0,d^2},$$

de sorte que

$$(S - 1/\alpha)\sqrt{n} \in \Phi_{0,d^2/\alpha^4}.$$

EXEMPLE 2. Trouver la distribution limite de la statistique

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

si $\mathbf{E}x_1 = \alpha$, $\mathbf{V}x_1 = d^2$ et $\mathbf{E}x_1^4 < \infty$. (Nous savons déjà que $S^2 \underset{\text{p.s.}}{\rightarrow} d^2$ en vertu du premier théorème de continuité.) On peut déterminer directement la distribution limite à l'aide des représentations

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha^2) - (\bar{x} - \alpha)^2,$$

$$(S^2 - d^2)\sqrt{n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(x_i - \alpha)^2 - d^2] - \sqrt{n}(\bar{x} - \alpha)^2.$$

Mais nous allons nous servir du théorème 1A. Aux termes de ce dernier nous devons poser

$$G(F) = \int (x - \alpha)^2\, dF(x) - \left(\int x\, dF(x) - \alpha\right)^2,$$

si bien que $g_1(x) = (x - \alpha)^2$, $g_2(x) = x$, $h(t) = t_1 - (t_2 - \alpha)^2$. Comme

$$\frac{\partial h(a)}{\partial t_1} = 1, \quad \frac{\partial h(a)}{\partial t_2} = 0$$

au point $a = (d^2, \alpha)$, il vient

$$(S^2 - d^2)\sqrt{n} \Rightarrow \xi, \quad \xi \in \Phi_{0,\nu^2}, \quad \nu^2 = \mathbf{E}(x_1 - \alpha)^4 - d^4.$$

EXEMPLE 3. *Statistique* χ^2. En conclusion de ce paragraphe on se propose d'étudier une statistique qui se rapporte aussi bien au premier qu'au deuxième type.

Considérons les statistiques construites à l'aide des fonctionnelles de la forme

$$G(F) = h\left(\int g\, dF\right), \tag{2}$$

où g est une fonction à variations bornées sur un intervalle $[a, b]$ tel que $F(a) = 0$ et $F(b) = 1$ (a et b peuvent être infinis). Puisque $\int g\,dF = g(b) - \int F\,dg$, la fonctionnelle $G(F)$ sera continue pour la métrique uniforme si seulement la fonction h est continue. Il est clair que cette classe de statistiques est l'intersection des classes des statistiques du premier et du deuxième type.

Ceci est valable aussi pour le cas où g est une fonction vectorielle de composantes g_i à variations bornées.

Considérons maintenant la partition de l'axe réel (l'espace $\mathscr{X}$) en intervalles disjoints $\Delta_1, \ldots, \Delta_r$ et posons $\nu_i = n\mathbf{P}_n^*(\Delta_i)$, $p_i = \mathbf{P}_0(\Delta_i)$ ($\mathbf{P}_0$ est la distribution associée à F_0, de sorte que $X \in \mathbf{P}_0$). On appelle *statistique* $\chi^2 = \chi^2(X)$, la statistique

$$\chi^2(X) = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i}.$$

Il est évident que c'est une statistique du deuxième type, car elle est associée, au facteur multiplicatif n près, à la fonctionnelle

$$G(F) = G_1(\mathbf{P}) = \sum_{i=1}^{r} \frac{(\mathbf{P}(\Delta_i) - \mathbf{P}_0(\Delta_i))^2}{\mathbf{P}_0(\Delta_i)}.$$

Pour représenter $\chi^2(X)$ comme une statistique du premier type, considérons une fonctionnelle (2)

$$G(F) = h\left(\int g\,d(F - F_0)\right),$$

où $h(u) = \sum_{j=1}^{r} u_j^2$ et la fonction vectorielle g a pour composantes

$$g_j(x) = \begin{cases} 1/\sqrt{p_j} & \text{si } x \in \Delta_j, \\ 0 & \text{sinon.} \end{cases}$$

Vu que la fonction h est dérivable, $\dfrac{\partial h(0)}{\partial u_j} = 0$, $\dfrac{\partial^2 h(0)}{\partial u_i \partial u_j} = 2\delta_{ij}$ (δ_{ij} est le symbole de Kronecker), en posant $S_n(X) = G(F_n^*)$, on obtient

$$nS_n(X) = n\sum_{j=1}^{r}\left[\left(\frac{\nu_j}{n} - p_j\right)\frac{1}{\sqrt{p_j}}\right]^2 = \chi^2(X).$$

Si $X \in \mathbf{P}_0$, il résulte de la deuxième partie du théorème 1A que

$$\chi^2(X) \Rightarrow \sum_{j=1}^{r} \xi_j^2, \tag{3}$$

où $\xi = (\xi_1, \ldots, \xi_r)$ est un vecteur $\left(\text{limite pour } \left(\frac{\nu_1 - np_1}{\sqrt{np_1}}, \ldots \right.\right.$ $\left.\left.\ldots, \frac{\nu_r - np_r}{\sqrt{np_r}}\right)\right)$ distribué suivant la loi normale, de moyenne nulle et de matrice $\sigma^2 = \|\sigma_{ij}\|$ des moments d'ordre deux,

$$\sigma_{ij} = \mathbf{E}\xi_i\xi_j = \mathbf{E}(g_i(x_1) - \sqrt{p_i})(g_j(x_1) - \sqrt{p_j})$$

(de la définition de g_j, il s'ensuit que $\mathbf{E}g_j(x_1) = \sqrt{p_j}$). Comme $g_i(x)g_j(x) = 0$ pour $i \neq j$, et $\mathbf{P}(g_j^2(x_1) = 1/p_j) = p_j$, $\mathbf{P}(g_j^2(x_1) = 0) = 1 - p_j$, il vient

$$\sigma_{ij} = \delta_{ij} - \sqrt{p_i p_j}\,.$$

Voyons maintenant de quelle forme est la distribution du second membre de (3) (c'est-à-dire la distribution limite de $\chi^2(X)$).

Considérons une transformation orthogonale de R^r de matrice C et le vecteur

$$\eta = \xi C.$$

Le vecteur η est, comme le vecteur ξ, distribué suivant la loi normale. En effet, dire que ξ est une variable aléatoire normale, revient à dire que sa fonction caractéristique est égale à (cf. [11])

$$\mathbf{E}e^{it\xi^T} = e^{-\frac{1}{2}t\sigma^2 t^T},$$

où $\sigma^2 = \|\sigma_{ij}\|$ est la matrice des moments d'ordre deux. Or la fonction caractéristique de η :

$$\mathbf{E}e^{it\eta^T} = \mathbf{E}e^{itC^T\xi^T} = e^{-\frac{1}{2}tC^T\sigma^2 Ct^T}$$

est de la même forme, donc η est un vecteur normal, mais de matrice des moments d'ordre deux $d^2 = C^T\sigma^2 C = \|d_{ij}\|$, de sorte que

$$d_{ij} = \mathbf{E}\eta_i\eta_j = \sum_{k,l} c_{li}\sigma_{lk}c_{kj} = \sum_{k,l} c_{li}(\delta_{lk} - \sqrt{p_l p_k})c_{kj} =$$
$$= \sum_l c_{li}c_{lj} - \Big(\sum_l c_{li}\sqrt{p_l}\Big)\Big(\sum_k c_{kj}\sqrt{p_k}\Big). \qquad (4)$$

Choisissons maintenant la matrice C de telle sorte que sa première colonne soit composée des coordonnées $c_{l1} = \sqrt{p_l}$ (ceci revient à fixer le premier vecteur du système de coordonnées image, chose possible puisque $\sum_{l=1}^{r} c_{l1}^2 = \sum p_l = 1$). Il est alors évident que le second terme de (4) est, en vertu de l'orthogonalité de C, égal à 1 si $i = j = 1$ et à 0 sinon. Ce qui signifie que $d_{11} = \mathbf{E}\eta_1^2 = 0$, $d_{ij} = \mathbf{E}\eta_i\eta_j = \delta_{ij}$ pour $i \geqslant 2$ et donc que η_1 est presque sûrement nul et les variables $\eta_2, \ldots, \eta_r$, indépendantes normales de paramètres

(0, 1). La matrice C étant orthogonale, il vient

$$\sum_{j=1}^{r} \xi_j^2 = \sum_{j=1}^{r} \eta_j^2 = \sum_{j=2}^{r} \eta_j^2,$$

$$\chi^2(X) \Rightarrow \sum_{j=2}^{r} \eta_j^2. \tag{5}$$

Le second membre de (5) suit une loi appelée *loi du* χ^2 à $r - 1$ degrés de liberté (cf. [11] ainsi que le § 2.2). Nous aurons souvent affaire à cette loi dans la suite.

La relation (5) sera prouvée au paragraphe suivant, ainsi qu'au § 3.16 par des considérations plus générales.

Dans les chapitres ultérieurs, on trouvera d'autres exemples d'application des théorèmes 1 et 1A.

§ 8*. Distribution limite des statistiques du deuxième type

On se bornera au cas où $\mathscr{X} = R$. La fonctionnelle $G(F_n^*)$ sera une variable aléatoire si c'est une application mesurable de $(D(-\infty, \infty), \sigma_D)$ dans $(R, \mathfrak{B})$. Mais dans la suite il nous sera plus commode d'étudier des fonctionnelles définies non pas sur $D(-\infty, \infty)$ mais sur $D(0, 1)$ (comparer avec le § 6).

A cet effet, construisons une application de $D(-\infty, \infty)$ dans $D(0, 1)$. Supposons que la fonction de répartition F_0 associée à l'échantillon est continue et monotone, ce qui assure l'existence de la fonction inverse $F_0^{-1}(t)$ (qui est égale au quantile d'ordre t de F_0). Il nous suffit de considérer les valeurs de $G(F)$ pour des fonctions F dont le support est contenu dans celui de F_0. A chaque fonction F associons la fonction

$$\tilde{F}(t) = F(F_0^{-1}(t)) \equiv FF_0^{-1}(t).$$

Il est évident que $N_{\tilde{F}} \subseteq [0, 1]$, où $N_{\tilde{F}}$ est le support de $\tilde{F}$, de sorte que $\tilde{F} \in D(0, 1)$ est une fonction de répartition. L'application réciproque de $D(0, 1)$ dans $D(-\infty, \infty)$ est définie par

$$F(u) = \tilde{F}(F_0(u)) \equiv \tilde{F}F_0(u).$$

Associons maintenant à la fonctionnelle G la fonctionnelle $\tilde{G}$ définie sur les fonctions de répartition $H \in D(0, 1)$ $(N_H \subseteq [0, 1])$ par l'égalité

$$\tilde{G}(H) = G(HF_0). \tag{1}$$

L'inversion de cette formule nous donne

$$G(F) = \tilde{G}(FF_0^{-1}).$$

Ces égalités ramènent l'étude des fonctionnelles $G(F)$ à celle des fonctionnelles $\tilde{G}(H)$ définies sur les fonctions de répartition de $D(0, 1)$. Ces égalités

entraînent

$$G(F_n^*) = \tilde{G}(F_n^* F_0^{-1}) = \tilde{G}(D_n^*), \qquad (2)$$

où la fonction

$$D_n^* = F_n^* F_0^{-1} \qquad (3)$$

n'est autre que la fonction de répartition empirique d'un échantillon issu d'une distribution uniforme sur [0, 1]. En effet, en vertu du théorème 6.1, le processus $nD_n^*(t) = nF_n^*(F_0^{-1}(t))$ admet la même loi de probabilité que le processus de Poisson $\pi(F_0(F_0^{-1}(t))) = \pi(t)$, $t \in [0, 1]$ (de paramètre $\lambda > 0$), sachant que $\pi(1) = n$. Ce qui nous donne, toujours en vertu du théorème 6.1, la proposition annoncée.

De ce qui précède il s'ensuit que l'étude de $G(F_n^*)$ se ramène à celle de la fonctionnelle $\tilde{G}$ de la fonction de répartition empirique d'une distribution uniforme sur [0, 1].

EXEMPLE 1. Soit $G(F) = \zeta_p$ le quantile d'ordre p d'une fonction de répartition F. Alors $\tilde{G}(H) = G(HF_0)$ est le quantile d'ordre p de la fonction de répartition HF_0 ou, ce qui revient au même (dans l'hypothèse où, pour simplifier, H est continue), la solution de l'équation $H(F_0(t)) = p$, soit $F_0^{-1}(H^{-1}(p))$.

Ceci exprime que le quantile empirique $\zeta_n^* = G(F_n^*) = \tilde{G}(D_n^*)$ (cf. (2) et (3)) de l'échantillon $X \in F_0$ n'est autre que la valeur de la fonction F_0^{-1} du quantile empirique $\eta_p^* = (D_n^*)^{-1}(p)$ d'ordre p d'un échantillon Y de distribution uniforme.

Si l'on réussit donc à trouver la distribution limite de η_p^*, on pourra déduire celle de ζ_p^* grâce aux théorèmes de continuité.

EXEMPLE 2. Considérons la fonctionnelle $G(F) = \sup_{-\infty < t < \infty} |F(t) - F_0(t)|$. Dans ce cas

$$\tilde{G}(H) = G(HF_0) = \sup_{-\infty < t < \infty} |H(F_0(t)) - F_0(t)| = \sup_{u \in [0, 1]} |H(u) - u|,$$

de sorte que

$$G(F_n^*) = \tilde{G}(D_n^*) = \sup_{u \in [0, 1]} |D_n^*(u) - u|,$$

et aux termes du § 6, la distribution de la statistique $G(F_n^*)$ ne dépendra pas de F_0 si F_0 est continue. De ce point de vue, on peut dire que la statistique $G(F_n^*)$ est invariante par rapport à une distribution continue de l'échantillon.

EXEMPLE 3. La fonctionnelle

$$G(F) = \int_{-\infty}^{\infty} |F(t) - F_0(t)|^k dF_0(t)$$

engendre aussi une statistique $G(F_n^*)$ invariante par F_0, puisque

$$\tilde{G}(H) = \int_0^1 |H(u) - u|^k du, \quad G(F_n^*) = \int_0^1 |D_n^*(u) - u|^k du.$$

EXEMPLE 4. Considérons la fonctionnelle

$$G(F) = \sum_{j=1}^{r} \frac{(\Delta_j F - \Delta_j F_0)^2}{\Delta_j F_0},$$

où $\Delta_j F$ sont les accroissements de la fonction F sur les intervalles $\Delta_j = [t_j, t_{j+1}[$ de partition de la droite réelle. Il est évident que $nG(F_n^*)$ n'est autre que la statistique χ^2 traitée dans l'exemple 7.3 comme une statistique du premier type.

On a

$$\tilde{G}(H) = G(HF_0) = \sum_{j=1}^{r} \frac{(\Delta_j HF_0 - \Delta_j F_0)^2}{\Delta_j F_0},$$

où

$$\Delta_j HF_0 = H(F_0(t_{j+1})) - H(F_0(t_j)) = \delta_j H,$$

$\delta_j H$ sont les accroissements de H sur les intervalles $\delta_j = [\tau_j, \tau_{j+1}[$, avec $\tau_j = F_0(t_j)$. En désignant la longueur d'un intervalle δ_j par la même lettre δ_j, on obtient donc

$$G(F_n^*) = \tilde{G}(F_n^* F_0) = \tilde{G}(D_n^*) = \sum_{j=1}^{r} (\delta_j D_n^* - \delta_j)^2/\delta_j.$$

Le dernier membre est la statistique $n^{-1}\chi^2$ construite au vu d'un échantillon Y de distribution uniforme avec la partition $\{\delta_j\}$. Ceci exprime en particulier que dans l'exemple 3 du paragraphe précédent on aurait pu se borner à l'étude d'une fonction de répartition uniforme F_0, bien que la statistique χ^2 ne soit pas invariante par F_0.

Sans restreindre la généralité on peut donc admettre que la fonctionnelle $G(F)$ est définie sur $D(0, 1)$ et que $F_0(t) = t$, $t \in [0, 1]$. Le passage aux fonctionnelles « primitives » qui se réalise à l'aide des formules (1) et (2) sera illustré dans des exemples ultérieurs.

Pour pouvoir déterminer la distribution limite des fonctionnelles du deuxième type $G(F_n^*)$, il faut, comme dans le paragraphe précédent, assujettir les fonctionnelles à des conditions de régularité.

Posons pour simplifier $\|x\| = \sup_{0 \le t \le 1} |x(t)|$.

DÉFINITION 1. On dit qu'une fonctionnelle $G(F)$ est *k fois continûment dérivable en un point* F_0 s'il existe une fonctionnelle $g(F_0, v)$ qui pour toute

fonction $v \in C(0, 1)$ et toute suite de fonctions $v_n \in D(0, 1)$ telle que $\| v_h - v \| \to 0$ avec h, vérifie les relations

$$\frac{G(F_0 + hv_h) - G(F_0)}{h^k} \to g(F_0, v),$$
$$g(F_0, v_h) \to g(F_0, v). \tag{4}$$

La dernière relation exprime de toute évidence que la fonctionnelle $g(F_0, v)$ appelée ***dérivée de G d'ordre k dans la direction de v***, est continue pour une métrique uniforme dans $C(0, 1)$.

REMARQUE 1. On rappelle que par F_0 on entendra partout une fonction de répartition uniforme sur $[0, 1]$.

Montrons que dans l'exemple 1, la fonctionnelle $G(F) = F^{-1}(p)$ est continûment dérivable au « point » $F_0(t) = t$, $t \in [0, 1]$.

En effet, par définition

$$G(F_0 + hv_h) = \max\{t : F_0(t) + hv_h(t) \leqslant p\}.$$

Cette fonctionnelle étant continue pour la métrique uniforme au point F_0, on peut poser $G(F_0 + hv_h) = p + \delta$, où $\delta = \delta(h) \to 0$ pour $h \to 0$. Par ailleurs, de la relation $\| v_h - v \| \to 0$, $v \in C(0, 1)$, il s'ensuit que $| v_h(p + \delta) - v_h(p) | \equiv r(h) \to 0$ avec h. Comme $F_0(p + \delta) = p + \delta$, pour $t = G(F_0 + pv_h) = p + \delta$ on obtient

$$F_0(t) + hv_h(t) = p + \delta + hv_h(p + \delta) = p + \delta + h(v_h(p) + \tau r(h)) \leqslant p,$$

où $|\tau| \leqslant 1$. On obtiendrait l'inégalité contraire en se servant du fait que $F_0(t + 0) + hv_h(h + 0) \geqslant p$. D'où il résulte que $\delta = -h(v_h(p) + \tau_1 r(h))$, $|\tau_1| \leqslant 1$, de sorte que

$$\frac{G(F_0 + hv) - G(F_0)}{h} = \frac{\delta}{h} \to -v(p).$$

La dérivée $g(F_0, v)$ est donc égale à

$$g(F_0, v) = -v(p). \blacktriangleleft \tag{5}$$

Dans l'exemple 2, la fonctionnelle $G(F) = \sup_{t \in [0, 1]} | F(t) - F_0(t) |$ est, de toute évidence, aussi continûment dérivable suivant n'importe quelle direction, puisque $G(F_0) = 0$,

$$g(F_0, v) = \frac{G(F_0 + hv)}{h} = \sup_{t \in [0, 1]} | v(t) | .$$

Dans l'exemple 3, la fonctionnelle $G(F) = \int_0^1 | F(t) - F_0(t) |^k dR(t)$, où $R(t)$ est une fonction quelconque à variations bornées, est (k fois) continûment dérivable suivant n'importe quelle direction, puisque

$$g(F_0, v) = \frac{G(F_0 + hv)}{h^k} = \int_0^1 | v(t) |^k dR(t).$$

Idem pour la fonctionnelle de l'exemple 4

$$G(F) = \sum_{j=1}^{r} \frac{(\Delta_j F - \Delta_j F_0)^2}{\Delta_j F_0}$$

qui est deux fois continûment dérivable, puisque

$$g(F_0, v) = \frac{G(F_0 + hv)}{h^2} = \sum_{j=1}^{r} \frac{(\Delta_j v)^2}{\Delta_j F_0}.$$

Les généralisations des fonctionnelles des exemples 2, 3 et 4 sont les fonctionnelles de la forme $G(F) = G_1(F - F_0)$, où la fonctionnelle G_1 est homogène au sens que $G_1(hv) = h^k G(v)$. Il est évident que ces fonctionnelles seront toutes dérivables.

Formulons maintenant le théorème fondamental relatif aux fonctionnelles du deuxième type. Supposons comme toujours que $F_0(t) \equiv t$, $t \in [0, 1]$.

THÉORÈME 1. *Si $X \in F_0$ et $G(F)$ est une fonctionnelle (k fois) dérivable au sens de la définition* 1, *alors*

$$[G(F_n^*) - G(F_0)]n^{k/2} \Rightarrow g(F_0, w^\circ),$$

où w° est un pont brownien.

DÉMONSTRATION. On sait (cf. par exemple [5]) que les compacts de l'espace $C(0, 1)$ des fonctions continues muni d'une métrique uniforme se décrivent comme suit. A toute fonction $\varphi(\Delta) > 0$, $\varphi(\Delta) \to 0$ pour $\Delta \to 0$, et à un nombre $N > 0$ correspond le compact

$$K = K(\varphi, N) = \{y \in C(0, 1) : \omega_\Delta(y) \leqslant \varphi(\Delta), \ | y(0) | \leqslant N\},$$

où $\omega_\Delta(y)$ est le module de continuité de y :

$$\omega_\Delta(y) = \sup_{|t-u| \leq \Delta} | y(t) - y(u) | .$$

Désignons par K_h l'ensemble

$$K_h = \{y \in D(0, 1) : \omega_\Delta(y) \leqslant \varphi(\Delta) \textit{ pour tous les } \Delta \geqslant h \ ; \ |y(0)| \leqslant N\}.$$

On appellera les ensembles K_h « précompacts » (ce terme recouvre un autre sens en analyse fonctionnelle) engendrés par le compact K. Il est clair que $K_{h_1} \subset K_{h_2}$ pour $h_1 \leqslant h_2$, $\bigcap_{n=1}^{\infty} K_{1/n} = K$ et que $K_h \subset (K)^{\varphi(h)}$, où $(K)^\epsilon$ est un ϵ-voisinage de K.

Montrons maintenant que pour tout $\delta > 0$ donné, il existe un compact K (donc les précompacts K_h engendrés par K) et une suite $h_n \to 0$ pour $n \to \infty$, tels que

$$\lim_{n\to\infty} \sup \mathbf{P}(w^n \notin K_{h_n}) \leqslant \delta. \tag{6}$$

En effet, le théorème 6.3 nous dit que pour toute fonctionnelle f continue pour une métrique uniforme, on a $f(w^n) \Rightarrow f(w^\circ)$, où $w^n(t) = \sqrt{n}(F_n^*(t) - t)$, $t \in [0, 1]$. Comme $\omega_\Delta(y)$ est une telle fonctionnelle, il vient $\omega_\Delta(w^n) \Rightarrow \omega_\Delta(w^\circ)$. Or $\omega_\Delta(w^\circ) \underset{\text{p.s.}}{\to} 0$ avec Δ, puisque les trajectoires w° sont presque sûrement continues. Donc, pour ϵ et δ donnés et pour Δ assez petit, on a

$$\mathbf{P}(\omega_\Delta(w^\circ) > \epsilon) \leqslant \delta.$$

En admettant, sans restreindre la généralité, que ϵ est un point de continuité de la distribution $\omega_\Delta(w^\circ)$ on trouve

$$\lim_{n\to\infty}\sup \mathbf{P}(\omega_\Delta(w^n) > \epsilon) \leqslant \delta.$$

Soient maintenant $\epsilon_k \downarrow 0$ une suite et $\Delta_k \downarrow 0$ des nombres tels que

$$\lim_{n\to\infty} \sup \mathbf{P}(\omega_{\Delta_k}(w^n) > \epsilon_k) \leqslant \delta/2^{k+1}.$$

Formons la fonction $\varphi(\Delta) = \epsilon_k$ pour $\Delta \in [\Delta_{k+1}, \Delta_k[$. Il est clair que $\varphi(\Delta) \to 0$ pour $\Delta \to 0$ et l'on peut envisager les précompacts K_h construits à l'aide de la fonction φ. Pour tout $k < \infty$, on a alors

$$\lim_{n\to\infty} \sup \mathbf{P}(w^n \notin K_{\Delta_k}) \leqslant \lim_{n\to\infty} \sup \sum_{j=1}^{k+1} \mathbf{P}(\omega_{\Delta_j}(w^n) > \epsilon_j) \leqslant$$

$$\leqslant \sum_{j=1}^{k+1} \lim_{n\to\infty} \sup \mathbf{P}(\omega_{\Delta_j}(w^n) > \epsilon_j) \leqslant \delta/2$$

(cette inégalité peut être mise en défaut pour $k = \infty$). Cette relation exprime que pour tout δ il existe une suite $h_n \to 0$ pour $n \to \infty$ telle que soit réalisée (6). Considérons maintenant la quantité

$$[G(F_n^*) - G(F_0)]n^{k/2} = g(F_0, w^n) + H_n(w^n),$$

où $H_n(x) = [G(F_0 + x/\sqrt{n}) - G(F_0)]n^{k/2} - g(F_0, x)$. Puisque $g(F_0, w^n) \Rightarrow$ $\Rightarrow g(F_0, w^\circ)$ en vertu du théorème 6.3 et de la définition 1, il nous suffit de nous assurer que

$$H_n(w^n) \underset{P}{\rightarrow} 0. \tag{7}$$

Remarquons que pour tout compact $K \subset C(0, 1)$ et toute suite $h_n \rightarrow 0$ pour $n \rightarrow \infty$, on a

$$\sup_{\substack{x \in D(0,\ 1) \\ x \in (K)^{h_n}}} |H_h(x)| \rightarrow 0. \tag{8}$$

En admettant le contraire on arrive à établir l'existence d'une suite $x_n \in D(0, 1)$ telle que $\| x_n - x \| \rightarrow 0$, $x \in C(0, 1)$, $\limsup_{n \rightarrow \infty} |H_n(x_n)| > 0$, ce qui contredit la dérivabilité de G.

Les relations (6) et (8) entraînent

$$\mathbf{P}(|H_n(w^n)| > \epsilon) \leqslant \mathbf{P}(|H_n(w^n)| > \epsilon,\ w^n \in K_{h_n}) + \mathbf{P}(w^n \notin K_{h_n}),$$

$$\limsup_{n \rightarrow \infty} \mathbf{P}(|H_n(w^n)| > \epsilon) \leqslant \delta.$$

Ce qui prouve (7) et avec elle le théorème, puisque δ est arbitraire. ◄

Poursuivons l'étude d'exemples.

Soit η_p^* un quantile empirique d'ordre p pour un échantillon Y issu d'une distribution uniforme sur [0, 1]. La relation (5) et le théorème 1 nous donnent alors

$$(\eta_p^* - p)\sqrt{n} \Rightarrow -w^\circ(p) \underset{d}{=} w^\circ(p).$$

Nous avons établi que dans le cas général, lorsque F_0 est une fonction de répartition continue arbitraire, on a

$$\zeta_p^* = F_0^{-1}(\eta_p^*).$$

Si l'on applique maintenant le troisième théorème de continuité, on obtient le

COROLLAIRE 1. *Si* $X_n \in F_0$, F_0 *est continûment dérivable en* ζ_p *et* $f(\zeta_p) = F_0'(\zeta_p) > 0$, *alors*

$$(\zeta_p^* - \zeta_p)\sqrt{n} \Rightarrow w^\circ(p)/f(\zeta_p).$$

On remarquera que les conditions de ce corollaire expriment la dérivabilité continue de F_0^{-1} au point p :

$$(F_0^{-1}(p))' = \frac{1}{F_0'(F_0^{-1}(p))} = \frac{1}{f(\zeta_p)}.$$

Comme $\mathbf{E}w^\circ(p) = 0$, $\mathbf{V}w^\circ(p) = \mathbf{E}(w(p) - pw(1))^2 = \mathbf{E}(w(p)(1-p) + p(w(1) - w(p)))^2 = p(1-p)^2 + p^2(1-p) = p(1-p)$, l'assertion de ce corollaire devient

$$(\zeta_p^* - \zeta_p)\sqrt{n} \in \Phi_{0,\,\sigma^2},\ \sigma^2 = p(1-p)/f^2(\zeta_p). \blacktriangleleft$$

Dans l'exemple 2, la fonctionnelle $G(F) = \sup_{0 \le t \le 1} |F(t) - F_0(t)|$ est dérivable, donc en vertu du théorème 1

$$G(F_n^*)\sqrt{n} \Rightarrow \sup_{0 \le t \le 1} |w^\circ(t)|.$$

La distribution de $\eta = \sup_{0 \le t \le 1} |w^\circ(t)|$ a été explicitée dans [75] :

$$\mathbf{P}(\eta < z) = K(z) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2z^2}.$$

La fonction $K(z)$ s'appelle *fonction de Kolmogorov.*

Nous avons vu dans le cas général que lorsque F_0 est une fonction de répartition continue arbitraire, la distribution de la statistique

$$D(X) = \sup_t |F_n^*(t) - F_0(t)|$$

est la même pour le cas où $F_0(t) = t$, $t \in [0, 1]$. On obtient ainsi le

COROLLAIRE 2 (théorème de Kolmogorov). *Si $X \in F_0$ et F_0 est continue, alors*

$$\sqrt{n}\, D(X) \in K.$$

Cela exprime que le désaccord maximal $D(X)$ entre la fonction $F_n^*(t)$ et la fonction $F_0(t)$ est de l'ordre de $1/\sqrt{n}$ et peut être approximativement mis sous la forme $D(X) \approx \eta/\sqrt{n}$.

Dans l'exemple 3, nous avons vu que l'autre statistique (qui est souvent désignée par ω^2)

$$\omega^2 = \int_{-\infty}^{\infty} (F_n^*(t) - F_0(t))^2 dF_0(t)$$

est aussi invariante par F_0. Le théorème 1 entraîne le

COROLLAIRE 3. *Si $X \in F_0$ et F_0 est continue, alors*

$$n\omega^2 \Rightarrow \int_0^1 [w^\circ(t)]^2 dt.$$

La distribution $\int_0^1 [w^\circ(t)]^2 dt$ a aussi été explicitée et tabulée au même titre que $K(z)$. Appliqué à l'exemple 4, le théorème 1 nous donne le

COROLLAIRE 4. *Si $X \in F_0$ et F_0 est continue, alors*

$$\chi^2 \Rightarrow \sum_{j=1}^{r} (\delta_j w^\circ)^2/\delta_j,$$

où δ_j, $j = 1, 2, \ldots, r$, représentent la partition de l'intervalle [0, 1] *définie dans l'exemple* 4.

Si l'on pose $\xi = (\xi_1, \ldots, \xi_r)$, $\xi_j = \delta_j w^\circ/\sqrt{\delta_j}$ et que l'on se serve du fait que $\delta_j w^\circ = \delta_j w - w(1)\delta_j$, où w est un processus wienérien standard, on obtient:

$$\chi^2 \Rightarrow \sum_{j=1}^{r} \xi_j^2, \quad \xi \in \Phi_{0,\,\sigma^2},$$

où la matrice $\sigma^2 = \|\sigma_{ij}\|$ est la même que dans l'exemple 7.3 puisque

$$\delta_j w^\circ = \delta_j w - \Big(\sum_k \delta_k w\Big)\delta_j = \sum_{k=1}^{r} a_{kj}\delta_k w,$$

$$a_{kj} = \delta_{kj} - \delta_j, \quad \mathsf{E}(\delta_k w)\,(\delta_l w) = \delta_{kl}\delta_k,$$

$$\sigma_{ij} = \frac{\mathsf{E}(\delta_i w^\circ)(\delta_j w^\circ)}{\sqrt{\delta_i\delta_j}} = \frac{1}{\sqrt{\delta_i\delta_j}} \sum_{k=1}^{r} a_{ki}a_{kj}\delta_k =$$

$$= \frac{1}{\sqrt{\delta_i\delta_j}}(\delta_{ij}\delta_i - \delta_i\delta_j) = \delta_{ij} - \sqrt{\delta_i\delta_j}$$

(δ_{kl} est le symbole de Kronecker). En reprenant les raisonnements de l'exemple 7.3, on trouve que $\sum_{j=1}^{r} \xi_j^2$ suit une loi du χ^2 à $r - 1$ degrés de liberté.

Signalons en conclusion de ce paragraphe que les statistiques qui présentent de l'intérêt ne sont pas censées être du premier ou du deuxième type. Il n'est qu'à citer la statistique $S(X) = \sum_{i=1}^{n-1} x_i x_{i+1}$ ou les statistiques S liées aux fonctionnelles $G_n(F)$, où G_n dépendent « essentiellement » de n (pas uniquement à travers l'échantillon), tel le terme maximal de l'échantillon ordonné $S(X) = x_{(n)} = \zeta^*_{1-1/n}$, etc.

§ 9*. Remarques sur les statistiques non paramétriques

La statistique ζ_p^* de l'exemple 8.1 se distingue essentiellement des statistiques des exemples 8.2, 8.3 et 8.4 par le fait que sa distribution limite est liée à la fonction de répartition F_0 (comparer avec le corollaire 8.1).

DÉFINITION 1. On dit qu'une statistique $S(X)$ est *asymptotiquement non paramétrique* si $S(X) \Leftarrow \mathbf{Q}$ lorsque $n \to \infty$, et $\mathbf{Q}$ ne dépend pas de la distribution de X, c'est-à-dire ne dépend pas de F_0 si $X \in F_0$.

Signalons que la fonction S peut fort bien dépendre de F_0. Le terme « non paramétrique » n'est pas très heureux ; il est cependant passé dans l'usage (son emploi est justifié lorsque la fonction F_0 appartient à une famille paramétrique : la distribution $\mathbf{Q}$ ne dépend pas d'un paramètre et en ce sens est non paramétrique). On se servira parfois du terme anglais *distribution free*.

Nous avons vu aux §§ 6, 7 et 8 que les statistiques $\sqrt{n}\, U(X)$, $\sqrt{n}\, D(X)$, $n\omega^2(X)$ et $\chi^2(X)$ sont asymptotiquement non paramétriques.

Le théorème 6.1 nous permet maintenant d'introduire une notion plus étroite. Dans ce théorème nous avons établi que $nF_n^*(t)$ est distribuée comme $\eta(F_0(t))$, où $\eta(u)$ est un processus poissonnien conditionnel de paramètre arbitraire $\lambda > 0$ sachant que $\eta(1) = n$ (cf. § 6), c'est-à-dire un processus indépendant de F_0. Si donc la statistique S est construite comme une fonctionnelle $G(F_n^*)$ (ou $G(F_n^* - F_0)$) invariante par un changement du « temps » t dans l'argument, sa distribution sera indépendante de F_0. Exemple :

$$D = \sup_t |F_n^*(t) - F_0(t)| \underset{d}{=} \frac{1}{n} \sup_t |\eta(F_0(t)) - nF_0(t)| =$$

$$= \frac{1}{n} \sup_{u \in [0,1]} |\eta(u) - un|. \qquad (1)$$

Ce qui précède nous inspire la

DÉFINITION 2. On dit qu'une statistique $S(X)$ est *non paramétrique* si sa distribution est indépendante de F_0 ($X \in F_0$).

Les relations (1) expriment que la statistique D est non paramétrique.

Nous avons signalé également (cf. corollaire 8.3) que la statistique ω^2, tout comme D, ne dépend pas de F_0, donc est aussi non paramétrique.

Etant asymptotiquement non paramétrique, la statistique χ^2 ne sera pas non paramétrique. On peut s'en assurer directement sur un exemple dans lequel on posera $r = 2$ et $n = 1$.

On obtient d'autres exemples de statistiques non paramétriques en considérant les valeurs $F_n^*(\zeta_p)$, où ζ_p est le quantile d'ordre p, de sorte que $nF_n^*(\zeta_p) \underset{d}{=} \eta(p)$ (cf. § 6). Le nombre r_j d'éléments de l'échantillon X inférieurs à x_j (ce nombre est dit *statistique de rang*) est aussi une statistique non paramétrique.

Les statistiques non paramétrique et asymptotiquement non paramétrique sont très utiles en théorie des tests des hypothèses statistiques (voir chapitre 3), puisque leurs distributions, qui sont nécessaires à la construction des tests, il suffit de les calculer une seule fois (par exemple pour une fonction de répartition uniforme F_0) et de les appliquer ensuite à toutes les autres distributions de l'échantillon.

§ 10*. Distributions empiriques lissées. Densités empiriques

Au § 2 nous avons associé à chaque échantillon X une distribution $\mathbf{P}_n^*$ que nous avons appelée empirique et qui est la somme de n distributions concentrées aux points $x_1, \ldots, x_n$. Cette distribution jouit de remarquables propriétés qui ont été décrites dans les paragraphes précédents. Mais la façon dont nous avons défini $\mathbf{P}_n^*$ n'est pas la seule possible et, dans bien des cas, pas la plus naturelle. Il existe d'autres procédés de définition de $\mathbf{P}_n^*$ qui non seulement conservent les propriétés des distributions empiriques étudiées plus haut, mais en font apparaître de nouvelles.

On se bornera à discuter la nature des distributions placées aux points x_i. Dans la définition que nous avons donnée de $\mathbf{P}_n^*$, c'étaient des distributions dégénérées $\mathbf{I}_{x_i}(B)$, de sorte que

$$\mathbf{P}_n^*(B) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_{x_i}(B). \tag{1}$$

Dans ce cas la distribution empirique est singulière pour la mesure de Lebesgue et n'admet donc pas de densité. Ceci peut être gênant dans les cas où l'on sait *a priori* que la distribution initiale $\mathbf{P}$ possède une densité. Dans ces conditions, il serait souhaitable d'avoir affaire à une distribution empirique $\mathbf{P}_n^*$ régulière telle que $\mathbf{P}_n^* \to \mathbf{P}$ et $f_n^* \to f$, où f_n^* et f sont les densités respectives de $\mathbf{P}_n^*$ et $\mathbf{P}$, la convergence étant entendue à tous les sens définis ci-dessus.

Ceci se réalise sans peine de la manière suivante. Soit $\mathbf{Q}$ une distribution admettant une densité. Posons

$$\mathbf{P}_n^{**}(B) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Q}\left(\frac{B - x_i}{h_n}\right), \tag{2}$$

où $\dfrac{B - x}{h}$ est l'ensemble des points $y \in \mathscr{X}$ tels que $x + yh \in B$; $h_n \to 0$ pour $n \to \infty$.

Il est évident que $\mathbf{P}_n^{**}(B)$ n'est autre que la « somme moyenne » des distributions $\mathbf{Q}$ réduites aux dimensions h_n et placées aux points x_i. La définition (2) généralise (1). La formule (1) se déduit de (2) pour $\mathbf{Q} = \mathbf{I}_0$, puisque

$\mathbf{I}_{x_i}(B) = \mathbf{I}_0(B - x_i) = \mathbf{I}_0\left(\frac{B - x_i}{h_n}\right)$ pour toute suite $\{h_n\}$.

Signalons les propriétés suivantes de la distribution $\mathbf{P}_n^{**}$ que nous appellerons *distribution empirique lissée*.

1. La distribution $\mathbf{P}_n^{**}$ est le produit de convolution des distributions $\mathbf{P}_n^{*}$ et $\mathbf{Q}(B/h_n)$, et

$$\mathbf{P}_n(B) = \mathbf{E}\mathbf{P}_n^{**}(B) = \int \mathbf{Q}\left(\frac{B - y}{h_n}\right)\mathbf{P}(dy)$$

le produit de convolution des distributions $\mathbf{P}$ et $\mathbf{Q}(B/h_n)$. Autrement dit, $\mathbf{P}_n(B)$ est la distribution de la variable aléatoire $\xi + h_n\eta$, où $\xi \in \mathbf{P}$, $\eta \in \mathbf{Q}$. Les théorèmes de continuité entraînent que

$$\mathbf{P}_n \Rightarrow \mathbf{P} \quad \text{lorsque} \quad h_n \to 0. \tag{3}$$

Rappelons que la distribution $\mathbf{P}_n^{*}$ était justiciable de l'égalité

$$\mathbf{E}\mathbf{P}_n^{*} = \mathbf{P}.$$

2. Si la distribution $\mathbf{P}$ est absolument continue pour la mesure de Lebesgue, la distribution $\mathbf{P}_n^{**}$ satisfera des théorèmes analogues à celui de Glivenko-Cantelli. En effet, dans ce cas la convergence (3) équivaudra à la convergence uniforme des distributions sur tous les intervalles. En se bornant, par souci de simplicité, à la dimension un, on aura ($F_n^{**}(x)$, $F_n(x)$ et $Q(x)$ étant les fonctions de répartition respectivement des distributions $\mathbf{P}_n^{**}$, $\mathbf{P}_n$ et $\mathbf{Q}$)

$$F_n^{**}(x) - F(x) = \int Q\left(\frac{x - y}{h_n}\right)dF_n^{*}(y) - F(x) =$$

$$= -\int F_n^{*}(y)d_yQ\left(\frac{x - y}{h_n}\right) - F(x) = F_n(x) - F(x) -$$

$$-\int \Big(F_n^{*}(y) - F(y)\Big)d_yQ\left(\frac{x - y}{h_n}\right).$$

Comme déjà signalé, la différence $F_n(x) - F(x) \to 0$ uniformément en x, tandis que l'intégrale du dernier membre, elle, est $\leqslant \sup_y |F_n^{*}(y) - F(y)|$, quantité qui tend presque sûrement vers 0.

3. L'avantage de $\mathbf{P}_n^{**}$ sur $\mathbf{P}_n^{*}$, avantage qui du reste a motivé l'introduction de cette distribution, est qu'*elle admet la densité*

$$f_n(x) = \frac{1}{nh_n}\sum_{i=1}^{n} q\left(\frac{x - x_i}{h_n}\right) = \frac{1}{h_n}\int q\left(\frac{x - y}{h_n}\right)dF_n^{*}(y) \tag{4}$$

($q(x)$ est la densité de la distribution **Q**) qui pour tout x tend vers la densité $f(x)$ de **P** lorsque $n \to \infty$ et $h_n \to 0$.

Avant de passer à la démonstration de cette assertion, on remarquera que pour obtenir de bons résultats sur la convergence de $f_n^*(x)$ vers $f(x)$, il faut se servir de densités q régulières et bornées. Si par exemple on choisit des densités q qui ne soient pas bornées, l'estimation $f_n^*(x)$ de la densité régulière $f(x)$ sera mauvaise. Comme le choix de q se trouve à notre discrétion, nous pouvons admettre qu'au moins est réalisée la condition suivante :

$$d^2 = \int q^2(t)dt < \infty. \tag{5}$$

THÉORÈME 1. *Si q vérifie la condition* (5), *$f(x)$ est continue et bornée, $h_n \to 0$ pour $n \to \infty$, de telle sorte que $nh_n \to \infty$, alors*

$$f_n^*(x) = f_n(x) + \zeta_n(x)/\sqrt{nh_n}, \tag{6}$$

où $f_n(x)$ est une fonction non aléatoire

$$f_n(x) = \mathbf{E}f_n^*(x) = \mathbf{E}h_n^{-1}q\left(\frac{x - x_1}{h_n}\right) = \frac{1}{h_n}\int q\left(\frac{x - t}{h_n}\right)f(t)\,dt =$$

$$= \int q(z)f(x - zh_n)dz \to f(x) \tag{7}$$

pour $h_n \to 0$. Les variables aléatoires $\zeta_n(x)$ sont asymptotiquement normales, $\zeta_n(x) \in \Phi_{0,\,\sigma^2(x)}$, $\sigma^2(x) = f(x)d^2$.

DÉMONSTRATION. La somme de (4) est une somme de variables aléatoires indépendantes équidistribuées dans un schéma de séries et de plus $f_n(x) = \mathbf{E}f_n^*(x)$ admet la représentation (7). Posons

$$\xi_{k,n} = \frac{1}{\sqrt{nh_n}}\left[q\left(\frac{x - x_k}{h_n}\right) - h_n f_n(x)\right].$$

Alors

$$f_n^*(x) - f_n(x) = \frac{1}{\sqrt{nh_n}}\sum_{k=1}^{n}\xi_{k,n}, \quad \mathbf{E}\xi_{k,n} = 0,$$

$$\mathbf{E}\xi_{k,n}^2 = \frac{1}{n}\left[\mathbf{E}\frac{1}{h_n}q^2\left(\frac{x - x_k}{h_n}\right) - h_n f_n^2(x)\right],$$

$$\mathbf{E}\frac{1}{h_n}q^2\left(\frac{x - x_k}{h_n}\right) = \frac{1}{h_n}\int q^2\left(\frac{x - t}{h_n}\right)f(t)\,dt =$$

$$= \int q^2(z)f(x - zh_n)dz \to f(x)\int q^2(z)dz = f(x)d^2. \tag{8}$$

Donc, $\mathbf{E}\xi_{k,n}^2 \sim f(x)d^2/n$ si $f(x)>0$. La condition de Lindeberg s'écrit ici

$$n\mathbf{E}(\xi_{1,n}^2\,;\ |\xi_{1,n}| > \epsilon)\to 0 \tag{9}$$

pour $n\to\infty$ et quel que soit $\epsilon>0$. Comme $h_n f_n^2(x)\to 0$ et $n\xi_{1,n}^2 \leqslant 2(q^2((x - x_1)/h_n) + h_n f_n^2(x))$, pour que (9) soit réalisée il suffit que

$$\mathbf{E}\left(\frac{1}{h_n}\,q^2\left(\frac{x - x_1}{h_n}\right)\,;\ q\left(\frac{x - x_1}{h_n}\right) > \epsilon\sqrt{nh_n}\right)\to 0.$$

Cette relation est satisfaite, puisque son premier membre est égal à (comparer avec (8))

$$\int\limits_{q(z)>\epsilon\sqrt{nh_n}} q^2(z)f(x - zh_n)\,dz \leqslant c \int\limits_{q(z)>\epsilon\sqrt{nh_n}} q^2(z)\,dz\to 0.$$

Donc, la variable aléatoire $\zeta_n(x) = \sum_{k=1}^{n} \xi_{k,n}$ est justiciable du théorème limite central. Ce qui prouve le théorème 1. ◄

Dans ce problème il est naturel de se poser la question du choix optimal de h_n et de la fonction $q(t)$. La réponse à cette question dépend de la régularité de $f(x)$. En effet, supposons par exemple que $f(x)$ n'est strictement positive que sur un intervalle fini, est bicontinûment dérivable et que $\varphi = \int (f''(x))^2 dx$ est fixe. Supposons par ailleurs que $\int zq(z)dz = 0$ (c'est toujours le cas pour des $q(z)$ symétriques) et que $D^2 = \int z^2 q(z)dz<\infty$. Alors

$$f_n(x) = \int q(z)f(x - zh_n)dz =$$
$$= \int q(z)\left[f(x) - zh_n f'(x) + \frac{z^2h_n^2}{2}f''(x) + o(z^2h_n^2)\right]dz =$$
$$= f(x) + \frac{h_n^2 f''(x)}{2}\int z^2 q(z)dz + o(h_n^2).$$

On voit que

$$\begin{aligned} f_n^*(x) - f(x) &= \frac{D^2h_n^2 f''(x)}{2} + \frac{\zeta_n(x)}{\sqrt{nh_n}} + o(h_n^2), \\ \mathbf{E}[f_n^*(x) - f(x)]^2 &= \left(\frac{D^2h_n^2f''(x)}{2}\right)^2 + \frac{d^2f(x)}{nh_n} + o(h_n^4). \end{aligned} \tag{10}$$

La minimisation de la dernière expression par rapport à h_n et q nous donne, en vertu de la normalité asymptotique de $\zeta_n(x)$, la plus petite « dispersion » de $f_n^*(x)$ autour de $f(x)$. Mais les valeurs minimisantes de h_n et de q dépendront de x par l'intermédiaire des valeurs inconnues de $f(x)$ et $f''(x)$. Pour éliminer cet effet et obtenir une optimalité « en moyenne », il est naturel d'envisager l'intégrale

$$\int \mathbf{E}[f_n^*(x) - f(x)]^2 dx \tag{11}$$

dont la partie principale sera égale à $\left(\frac{D^2 h_n^2}{2}\right)^2 \varphi + \frac{d^2}{nh_n}$ (on obtient ceci en supprimant $o(h_n^4)$ dans (10)).

Le minimum de cette expression est atteint pour $h_n = \left(\frac{d^2}{nD^4\varphi}\right)^{1/5}$ L'intégrale (11) sera alors égale à

$$\frac{5}{4}\varphi^{1/5}(Dd^2)^{4/5}n^{-4/5} + o(n^{-4/5}), \tag{12}$$

$$f_n^*(x) - f(x) = \left(\frac{Dd^2}{n\varphi}\right)^{2/5}\left(\frac{f''(x)}{2} + f(x)\sqrt{\varphi}\,\xi_n\right) + o(n^{-2/5}),$$

$$\xi_n \in \Phi_{0,\,1}.$$

Donc, la vitesse de convergence est ici de l'ordre de $n^{-2/5}$ contrairement à celle des fonctions de répartition qui est de l'ordre de $n^{-1/2}$. Ce fait est logique, puisque dans l'estimation de la valeur $f(x)$ ne participent, *grosso modo*, que les observations qui sont concentrées dans un voisinage décroissant du point x.

L'expression (12) permet de choisir de façon optimale la fonction $q(z)$ aussi, c'est-à-dire la fonction qui minimise Dd^2. En admettant, sans nuire à la généralité, que $D = 1$, on obtient le problème de minimisation de $d^2 = \int q^2(z)\,dz$ sous les conditions $\int q(z)\,dz = \int z^2 q(z)\,dz = 1$, $\int zq(z)\,dz = 0$.

Signalons que si f admet des dérivées continues d'ordre $2m > 2$, on peut obtenir de plus grandes vitesses de convergence de la différence $f_n^*(x) - f(x)$ vers 0. Pour cela il faut se servir de distributions généralisées **Q** dont les « densités » q sont de signe + ou − et permettent de satisfaire les conditions $\int z^{2m} q(z)\,dz = 1$, $\int z^j q(z)\,dz = 0$ pour tous les $j \in [1, 2m-1]$. Dans ce cas, en reproduisant les mêmes raisonnements, on pourra obtenir une vitesse de convergence de l'ordre de $n^{-\frac{2m}{4m+1}} = n^{-1/2 + \frac{1}{2(4m+1)}}$ et qui sera d'autant meilleure que m sera grand. Cette circonstance s'explique par le fait que dans l'estimation des valeurs $f(x)$ de fonctions $f(x)$ plus régulières participent les éléments de l'échantillon qui sont situés dans des voisinages plus vastes du point x.

D'autre part, on peut choisir les fonctions régulières $q(z)$ de telle sorte qu'il soit possible d'estimer et la densité $f(x)$ et ses dérivées. On peut s'en assurer aussi à l'aide des raisonnements produits ci-dessus.

Les fonctions $f_n^*(x)$ de forme (4) sont souvent appelées *estimateurs de Rosenblatt-Parzen* de la densité $f(x)$, ou encore *estimateurs nucléaires* de $f(x)$. Les fonctions $q(z)$ sont dites *noyaux*. On se sert souvent en pratique

des noyaux « rectangulaires », c'est-à-dire on admet que

$$q(z) = \begin{cases} 1 & \text{si} \quad z \in [-1/2, 1/2] \\ 0 & \text{ailleurs.} \end{cases}$$

On procède parfois plus simplement : on partage la droite réelle en petits intervalles Δ_j (de longueur h_n) et on pose $f_n^*(x) = \frac{\nu_j}{nh_n}$ pour $x \in \Delta_j$, où ν_j est le nombre d'éléments de l'échantillon contenus dans Δ_j. La fonction $f_n^*(x)$ s'appelle *histogramme* de l'échantillon. On vérifie immédiatement que si $f(x)$ est continue, l'histogramme $f_n^*(x)$ convergera, comme la fonction (4), vers $f(x)$ en probabilité si seulement $h_n \to 0$ et $nh_n \to \infty$.

CHAPITRE 2

THÉORIE DE L'ESTIMATION DES PARAMÈTRES INCONNUS

Le § 2 passe en revue les familles paramétriques de distributions usuelles et leurs propriétés fondamentales.

Les §§ 3 à 6 développent les principales méthodes d'estimations ponctuelles.

Les §§ 7 et 8 discutent les diverses approches de comparaison des estimateurs.

Les §§ 9 à 20 traitent des méthodes de construction des estimateurs optimaux (dans tel ou tel sens). Quatre directions sont dégagées :

1) (§§ 9, 10, 11 et 20) Approches bayésienne et minimax de construction des estimateurs optimaux. Les §§ 9 et 10 sont accessoires et contiennent les définitions et les principales propriétés des espérances mathématiques conditionnelles et des distributions conditionnelles.

2) (§§ 12 à 15) Construction des estimateurs optimaux (efficaces) exhaustifs et sans biais.

3) (§§ 16, 17, 22) Construction des estimateurs optimaux (efficaces) à l'aide de l'inégalité de Rao-Cramer.

4) (§§ 18, 19) Utilisation de l'invariance.

Dans les §§ 21 à 29 on étudie les propriétés asymptotiques du rapport de vraisemblance. On applique ensuite ces propriétés pour établir l'optimalité asymptotique des estimateurs du maximum de vraisemblance. Les résultats des §§ 21 à 29 servent de base à la théorie des tests asymptotiquement optimaux, développée au chapitre 3.

Les §§ 31 et 32 sont consacrés aux estimations par intervalles.

§ 1. Remarques préliminaires

Nous avons déjà noté dans les paragraphes précédents que l'objet liminal des recherches statistiques est un échantillon

$$X_n = (x_1, \ldots, x_n), \quad x_i \in \mathscr{X},$$

issu d'une distribution **P** entièrement ou partiellement inconnue. La statistique mathématique traite traditionnellement deux classes de problèmes :

1. *L'estimation des paramètres inconnus.*

2. *Le test des hypothèses statistiques.*

Les problèmes de la première classe se posent lorsqu'il faut, au vu d'un échantillon $X = X_n$, estimer une caractéristique numérique inconnue θ de la distribution **P**. Autrement dit, étant donné une fonctionnelle

$$\theta = \theta(\mathbf{P})$$

de la distribution **P**, on demande une fonction d'échantillon (ou ce qui est équivalent une statistique)

$$\theta^* = \theta_n^*(X_n),$$

qui puisse être utilisée à la place du paramètre θ. Nous avons vu dans le chapitre précédent que cela était possible. La statistique θ^* s'appelle *estimateur* du paramètre θ. On conçoit aisément que le paramètre θ admet une multitude d'estimateurs. Le théorème 1.3.1 suggère par exemple de prendre pour estimateur de la fonctionnelle

$$\theta = \int g(x)\, dF(x)$$

la statistique

$$\theta^* = \frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

On pourrait certes envisager d'autres estimateurs, par exemple

$$\theta^* = \frac{1}{n - \nu_1 - \nu_2} \sum_{j=\nu_1+1}^{n-\nu_2} g(x_{(j)}),$$

où $x_{(j)}$, $j = 1, \ldots, n$, sont les éléments de l'échantillon ordonné, etc. Pour θ^* on peut prendre aussi des valeurs ne dépendant pas de l'échantillon. On peut même poser $\theta^* \equiv 0$, bien que cela ne soit pas toujours justifié et même mauvais si l'ensemble des valeurs possibles de θ ne contient pas la valeur 0.

Signalons au sujet de cette dernière remarque que souvent dans la position du problème d'estimation, on spécifie l'ensemble Θ des valeurs possibles du paramètre θ. Si, par exemple, l'on estime le taux θ de minéral contenu dans du minerai, il est évident que $\theta \in [0, 1]$.

Dans de nombreux cas, on sait à l'avance que la distribution **P** de l'échantillon X ne peut être arbitraire, mais appartient à une famille bien définie de distributions $\mathscr{P}$.

L'exemple 1 de l'Introduction est un problème d'estimation des paramètres.

Les problèmes de la deuxième classe portent sur le test de telle ou telle hypothèse concernant la distribution inconnue **P**. On peut par exemple éprouver l'hypothèse que **P** est d'une forme donnée. A ce type de problèmes se rapporte l'exemple 2 de l'Introduction.

Nous verrons ultérieurement qu'il n'y a pas de différence radicale de nature entre les problèmes de ces deux classes.

Dans ce chapitre, nous indiquons les positions des problèmes et les méthodes de résolution qui sont étroitement liées aux résultats du chapitre

précédent et que l'on pourrait qualifier de « purement statistiques » à la différence des méthodes générales de théorie des jeux qui seront développées à la fin de l'ouvrage (cf. avant-propos).

Les approches purement statistiques expriment, dans une certaine mesure, le principe des méthodes de statistique mathématique. Historiquement, elles ont pris forme bien avant les méthodes plus générales. Pour ce qui est de leur application, l'homme a dû probablement s'en servir explicitement ou implicitement tout au long de l'histoire de son savoir.

Tout ceci justifie l'exposition séparée des méthodes purement statistiques, bien que certains aspects de cet exposé puissent être considérés comme des cas particuliers de concepts plus généraux. Nous mettrons en même temps en évidence l'incapacité de l'approche purement statistique à poser des problèmes plus précis. Ceci nous permettra de comprendre l'adéquation des autres points de vue.

§ 2. Quelques familles paramétriques de distributions et leurs propriétés

Considérons quelques familles de distributions dépendant de paramètres (familles paramétriques de distributions) qui se présentent souvent dans les applications et qui apparaîtront ultérieurement soit dans le cadre de l'exposé, soit comme illustrations de cet exposé.

1. Distribution normale sur la droite. Par Φ_{α,σ^2} nous désignerons une distribution normale de paramètres (α, σ^2), c'est-à-dire une distribution de densité

$$\varphi_{\alpha,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}},$$

de sorte que

$$\Phi_{\alpha,\sigma^2}(B) = \int_B \varphi_{\alpha,\sigma^2}(x)\, dx.$$

Si $\xi \in \Phi_{0,1}$ et $k \geqslant 0$ est un entier, il est évident que

$$\mathbf{E}\xi^{2k+1} = 0.$$

En se servant du changement $x = \sqrt{2u}$, on trouve pour les moments d'ordre pair

$$\mathbf{E}\xi^{2k} = \frac{2}{\sqrt{2\pi}} \int_0^\infty x^{2k} e^{-x^2/2} dx = \frac{2^{k+1}}{\sqrt{2\pi}} \int_0^\infty u^k e^{-u} \frac{du}{\sqrt{2u}} = \frac{2^k}{\sqrt{\pi}} \Gamma(k + 1/2),$$

où

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1}e^{-x}dx \tag{1}$$

est la fonction gamma, $\Gamma(\lambda) = (\lambda - 1)\Gamma(\lambda - 1)$, $\Gamma(1/2) = \sqrt{\pi}$, de sorte que

$$\mathsf{E}\xi^{2k} = (2k-1)!! = (2k-1)(2k-3)\ldots 1.$$

On obtiendrait ce résultat en dérivant $2k$ fois la fonction caractéristique $e^{-t^2/2}$ au point $t = 0$.

2. Distribution normale multidimensionnelle. Pour $\mathscr{X} = R^m$, le symbole $\Phi_{\alpha,\,\sigma^2}$ désignera une distribution normale dans R^m d'espérance mathématique $\alpha = (\alpha_1, \ldots, \alpha_m)$ et de matrice des moments centrés d'ordre deux $\sigma^2 = \|\sigma_{ij}\|$, $i, j = 1, \ldots, m$. Si A, la matrice inverse de σ^2, existe, la densité $\varphi_{\alpha,\,\sigma^2}(x)$ de la distribution $\Phi_{\alpha,\,\sigma^2}$ sera de la forme (cf. [11])

$$\varphi_{\alpha,\,\sigma^2}(x) = \frac{\sqrt{|A|}}{(2\pi)^{m/2}}\exp\left(-\frac{1}{2}(x-\alpha)A(x-\alpha)^T\right),$$

où x^T est le vecteur transposé du vecteur x. On rappelle aussi (ce fait a déjà été utilisé au § 1.7) que la fonction caractéristique de la variable $\xi \in \Phi_{\alpha,\,\sigma^2}$ est égale à

$$\mathsf{E}e^{it\xi^T} = \exp\left(it\alpha^T - \frac{1}{2}t\sigma^2t^T\right),$$

où $t = (t_1, \ldots, t_m)$ est un vecteur de R^m.

3. Distribution gamma. Le symbole $\Gamma_{\alpha,\,\lambda}$ désignera une distribution gamma de paramètres (α, λ). La densité $\gamma_{\alpha,\,\lambda}(x)$ de cette distribution dépend des deux paramètres $\alpha > 0$ et $\lambda > 0$ et vaut (cf. [11], § 7, chap. 6)

$$\gamma_{\alpha,\,\lambda}(x) = \begin{cases} \dfrac{\alpha^\lambda}{\Gamma(\lambda)}x^{\lambda-1}e^{-\alpha x}, & x \geqslant 0, \\ 0, & x < 0, \end{cases} \tag{2}$$

où $\Gamma(\lambda)$ est la fonction gamma définie dans (1). La fonction caractéristique de la distribution gamma s'écrit ([11])

$$\int_0^\infty e^{itx}\gamma_{\alpha,\,\lambda}(x)dx = \left(1 - \frac{it}{\alpha}\right)^{-\lambda}. \tag{3}$$

Si $\xi \in \Gamma_{\alpha,\,\lambda}$, on a

$$\mathsf{E}\xi^t = \frac{\alpha^\lambda}{\Gamma(\lambda)}\int_0^\infty x^{\lambda+t-1}e^{-\alpha x}dx = \frac{\alpha^{-t}}{\Gamma(\lambda)}\int_0^\infty y^{\lambda+t-1}e^{-y}dy = \frac{\alpha^{-t}\Gamma(\lambda+t)}{\Gamma(\lambda)}. \tag{4}$$

On obtiendrait le même résultat pour les $t > 0$ entiers en dérivant la fonction caractéristique. En posant $t = 1, 2$, on trouve

$$\mathsf{E}\xi = \lambda/\alpha, \quad \mathsf{V}\xi = \lambda/\alpha^2. \tag{5}$$

On voit sur les formules (3) et (4) que le paramètre α joue le rôle d'un paramètre d'échelle (ou de dispersion), de sorte que

$$\eta/\alpha \in \Gamma_{\alpha,\lambda} \quad \text{si} \quad \eta \in \Gamma_{1,\lambda}.$$

Cette circonstance nous suggère d'étudier de nombreuses propriétés de la distribution gamma pour une valeur seulement de α, par exemple pour $\alpha = 1$ ou $\alpha = 1/2$. La deuxième valeur est plus intéressante, car la distribution $\Gamma_{1/2,\lambda}$ joue un rôle important en statistique mathématique et s'appelle *distribution* χ^2.

4. Distribution χ^2 à k degrés de liberté. C'est ainsi qu'on appelle la distribution $\mathbf{H}_k = \Gamma_{1/2,\,k/2}$ pour $k > 0$ entier. Nous conserverons cette dénomination de $\mathbf{H}_k$ pour $k > 0$ quelconque. La fonction caractéristique de la distribution $\mathbf{H}_k$ est égale en vertu de (3) à

$$(1 - 2it)^{-k/2}.$$

Signalons les trois propriétés suivantes de la distribution $\mathbf{H}_k$.

1) Si η_i sont des variables aléatoires indépendantes de distribution $\mathbf{H}_{k_i}$, $i = 1, \ldots, s$, alors

$$\sum_{i=1}^{s} \eta_i \in \mathbf{H}_k, \quad k = \sum_{i=1}^{s} k_i.$$

Cette propriété résulte directement de la forme de la fonction caractéristique de $\mathbf{H}_k$.

2) *Si $\xi \in \Phi_{\alpha,\,\sigma^2}$, où $\Phi_{\alpha,\,\sigma^2}$ est une distribution normale k-dimensionnelle de matrice des moments d'ordre deux non dégénérée, alors*

$$Q(\xi) = (\xi - \alpha)\sigma^{-2}(\xi - \alpha)^T \in \mathbf{H}_k.$$

En effet, la fonction caractéristique de la variable aléatoire $Q(\xi)$ est égale à

$$\mathsf{E}e^{itQ(\xi)} = \frac{\sqrt{|\sigma^{-2}|}}{(2\pi)^{k/2}} \int \exp\left(-\frac{1}{2}Q(x)(1 - 2it)\right) dx_1, \ldots dx_k.$$

En effectuant le changement de variables $x_j\sqrt{1 - 2it} = y_j$, on obtient l'expression

$$(1 - 2it)^{-k/2} \frac{\sqrt{|\sigma^{-2}|}}{(2\pi)^{k/2}} \int e^{-\frac{1}{2}Q(y)}\, dy_1 \ldots dy_k = (1 - 2it)^{-k/2},$$

c.q.f.d. L'indépendance de l'intégrale du premier membre par rapport au domaine d'intégration résulte de l'analyticité de l'intégrant et de sa décroissance rapide lorsque $|y| \to \infty$ (cf. [11]).

De ce qui précède il s'ensuit que la variable aléatoire

$$\chi^2 = \xi_1^2 + \ldots + \xi_k^2,$$

où ξ_j sont des variables aléatoires indépendantes normales réduites, admet la distribution $\mathbf{H}_k$. Le terme « nombre de degrés de liberté » est lié précisément à cette représentation.

3) *Comme* $\mathsf{E}\xi_1^2 = 1$, $\mathsf{E}\xi_1^4 = 3$, $\mathsf{V}\xi_1^2 = 2$ *pour* $\xi_1 \in \Phi_{0,\,1}$, *il s'ensuit en vertu du théorème limite central que pour* $k \to \infty$

$$\frac{\chi^2 - k}{\sqrt{2k}} \Subset \Phi_{0,\,1}. \tag{6}$$

Ceci et les théorèmes de continuité du § 1.5 entraînent

$$\sqrt{2\chi^2} - \sqrt{2k - 1} \Subset \Phi_{0,\,1}.$$

Cette convergence est à l'origine de l'égalité approchée (pour de grands k et x) : $\mathbf{H}_k(]0, x[) \approx \Phi(\sqrt{2x} - \sqrt{2k - 1})$, $\Phi(x) = \Phi_{0,\,1}(]-\infty, x[)$, qui en principe est plus précise que l'approximation $\mathbf{H}_k(]0, x[) \approx \Phi\left(\dfrac{x - k}{\sqrt{2k}}\right)$ qui résulte de (6).

Signalons encore un cas particulier de la distribution gamma qui est fréquent dans les applications.

5. Distribution exponentielle. C'est la distribution $\Gamma_{\alpha,\,1}$ de densité

$$\alpha e^{-\alpha x}, \quad x > 0.$$

Des formules (5), il s'ensuit que pour $\xi \in \Gamma_{\alpha,\,1}$ on a

$$\mathsf{E}\xi = 1/\alpha, \quad \mathsf{V}\xi = 1/\alpha^2.$$

Considérons maintenant quelques distributions rattachées aux distributions normale et gamma, qui sont d'une grande importance en statistique mathématique. C'est la première fois que nous avons affaire à ces distributions.

6. Distribution $\mathbf{F}_{k_1,\,k_2}$ de Fisher à (k_1, k_2) degrés de liberté. C'est ainsi qu'on appelle la distribution de la variable aléatoire

$$\zeta = \eta_1/\eta_2,$$

où $\eta_j \in \mathbf{H}_{k_j}$, $j = 1, 2$, et sont indépendantes. Des propriétés de la distribution gamma il s'ensuit que ζ admet la même distribution lorsque $\eta_j \in \Gamma_{\alpha,\,k_j/2}$ quel que soit $\alpha > 0$, et que pour les k_j entiers, la variable ζ se

représente par

$$\zeta = \frac{\xi_1^2 + \ldots + \xi_{k_1}^2}{\zeta_1^2 + \ldots + \zeta_{k_2}^2},$$

où ξ_j et ζ_k sont des variables aléatoires indépendantes normales réduites. Calculons la densité de la distribution $\mathbf{F}_{k_1,\, k_2}$. On a

$$\mathbf{P}(\zeta < x) = \iint\limits_{u/v<x} \Gamma_{1,\,\lambda_1}(du)\,\Gamma_{1,\,\lambda_2}(dv) = \int\limits_{v=0}^{\infty}\int\limits_{u=0}^{vx} \frac{u^{\lambda_1-1}v^{\lambda_2-1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, e^{-u-v}\,du\,dv\,;$$

$$f_{(\zeta)}(x) = \frac{d\mathbf{P}(\zeta < x)}{dx} = \int\limits_0^{\infty} \frac{(vx)^{\lambda_1-1}v^{\lambda_2-1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, e^{-v-vx}v\,dv =$$

$$= \frac{x^{\lambda_1-1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\int\limits_0^{\infty} v^{\lambda_1+\lambda_2-1}e^{-v(1+x)}dv = \frac{x^{\lambda_1-1}}{(1+x)^{\lambda_1+\lambda_2}}\,\frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}. \quad (7)$$

On obtient la densité cherchée en faisant $\lambda_j = k_j/2$. Les moments de ζ (s'ils existent) sont :

$$\mathbf{E}\zeta^l = \frac{\Gamma(\lambda_1+\lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\int\limits_0^{\infty}\frac{x^{\lambda_1+l-1}}{(1+x)^{\lambda_1+\lambda_2}}\,dx = \frac{\Gamma(\lambda_1+l)\Gamma(\lambda_2-l)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}. \quad (8)$$

En particulier, pour $l = 1, 2$, on trouve

$$\mathbf{E}\zeta = \frac{\lambda_1}{\lambda_2 - 1}, \quad \mathbf{E}\zeta^2 = \frac{\lambda_1(\lambda_1+1)}{(\lambda_2-1)(\lambda_2-2)}.$$

La distribution de Fisher est parfois appelée *distribution de Snedécor*. Ceci est lié au fait que Fisher a proposé d'utiliser et tabulé non pas la distribution de ζ, mais celle de $\frac{1}{2}\ln\zeta$. La distribution de ζ a été tabulée plus tard par Snedécor.

7. Distribution $\mathbf{T}_k$ de Student *) à k degrés de liberté. Par définition, c'est la distribution de la variable aléatoire

$$t = \frac{\xi_0}{\sqrt{\frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2)}},$$

*) Student est le nom de plume de Gosset W.

où $\xi_j \in \Phi_{0,1}$, $j = 0, \ldots, k$, et sont indépendantes. Il est évident que $-t$ admet la même distribution. Donc, la distribution $\mathbf{T}_k$ est symétrique par rapport à l'origine des coordonnées. D'autre part,

$$t^2 = \frac{k\xi_0^2}{\xi_1^2 + \ldots + \xi_k^2} = \frac{k\eta_1}{\eta_2},$$

où η_j sont indépendantes, $\eta_1 \in \mathbf{H}_1$, $\eta_2 \in \mathbf{H}_k$. Ceci exprime que t^2/k suit la distribution de Fisher. Considérons la variable aléatoire $\tau = \sqrt{\zeta}$, $\zeta = \eta_1/\eta_2$, $\eta_j \in \mathbf{H}_{k_j}$. Puisque $\mathbf{P}(\tau < x) = \mathbf{P}(\zeta < x^2)$, la densité $f_{(\tau)}(x)$ de τ est égale à

$$f_{(\tau)}(x) = 2xf_{(\zeta)}(x^2) = 2x\frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \cdot \frac{x^{2\lambda_1 - 2}}{(1 + x^2)^{\lambda_1 + \lambda_2}} =$$

$$= \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \cdot \frac{2x^{2\lambda_1 - 1}}{(1 + x^2)^{\lambda_1 + \lambda_2}}, \quad \lambda_j = k_j/2, \quad x > 0.$$

En faisant $\lambda_1 = 1/2$ et $\lambda_2 = k/2$, on obtient de toute évidence la densité de $|t|/\sqrt{k}$. La distribution de t étant symétrique, sa densité $f_{(t)}(x)$ est en définitive

$$f_{(t)}(x) = \frac{\Gamma((k + 1)/2)}{\sqrt{\pi k}\,\Gamma(k/2)}\left(1 + \frac{x^2}{k}\right)^{-(k+1)/2}. \tag{9}$$

Il est clair que tous les moments de t d'ordre impair (s'ils existent) sont nuls. Pour les moments d'ordre $2l$ on a en vertu de (8)

$$\mathbf{E}t^{2l} = k^l\mathbf{E}\zeta^l = k^l\frac{\Gamma(\lambda_1 + l)\Gamma(\lambda_2 - l)}{\Gamma(\lambda_1)\Gamma(\lambda_2)},$$

où $\lambda_1 = 1/2$, $\lambda_2 = k/2$, $2l < k$. Pour $l = 1$, on obtient

$$\mathbf{E}t^2 = \frac{k}{k - 2}.$$

La fonction $f_{(t)}(x)$ rappelle par sa forme la densité de la loi normale. Bien plus,

$$f_{(t)}(x) \to \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad k \to \infty,$$

ce qui exprime que $t \in \Phi_{0,1}$ lorsque $k \to \infty$. Mais $f_{(t)}(x)$ admet des « ailes plus lourdes », puisque la fonction (9) décroît, lorsque $|x|$ croît, bien plus lentement que $e^{-x^2/2}$, de sorte que pour tous les $b > 0$

$$\mathbf{T}_k(]-b, b[) < \Phi_{0,1}(]-b, b[). \tag{10}$$

Ceci étant, l'écart entre le premier et le second membre de (10) peut être considérable pour les petits k.

Le lecteur peut prouver la convergence de $t = \sqrt{k}\,\xi_0/\sqrt{\eta_2}$ d'une autre façon en mettant à profit les théorèmes de continuité. Il suffit par exemple de remarquer que $\frac{\eta_2}{k} = \frac{1}{k}(\xi_1^2 + \ldots + \xi_k^2) \underset{\text{p.s.}}{\to} 1$, donc que $t \underset{\text{p.s.}}{\to} \xi_0$, $t \Rightarrow \xi_0$.

8. Distribution bêta. On appelle ainsi la distribution $\mathbf{B}_{\lambda_1,\,\lambda_2}$ de densité

$$f_{(\beta)}(x) = \begin{cases} \dfrac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, x^{\lambda_1 - 1}(1 - x)^{\lambda_2 - 1}, & x \in [0, 1], \\ 0, & x \notin [0, 1]. \end{cases}$$

Cette distribution doit son nom à la fonction bêta

$$\mathrm{B}(\lambda_1, \lambda_2) = \int_0^1 x^{\lambda_1 - 1}(1 - x)^{\lambda_2 - 1}\, dx = \frac{\Gamma(\lambda_1)\Gamma(\lambda_2)}{\Gamma(\lambda_1 + \lambda_2)}.$$

La distribution bêta est liée aux distributions gamma et de Fisher par la proposition suivante :

Si $\eta_j \in \Gamma_{\alpha,\,\lambda_j}$ (ou $\eta_j \in \mathbf{H}_{2,\lambda_j}$) et sont indépendantes, alors

$$\beta = \frac{\eta_1}{\eta_1 + \eta_2} = \frac{\zeta}{\zeta + 1} \in \mathbf{B}_{\lambda_1,\,\lambda_2},$$

où $\zeta = \eta_1/\eta_2 \in \mathbf{F}_{2\lambda_1,\,2\lambda_2}$.

Ce fait se prouve sans difficultés, puisqu'en vertu de (7) $\mathbf{P}(\beta < x) = \mathbf{P}\left(\zeta < \frac{x}{1 - x}\right)$,

$$f_{(\beta)}(x) = f_{(\zeta)}\left(\frac{x}{1 - x}\right)\left(\frac{x}{1 - x}\right)' = \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\left(\frac{x}{1 - x}\right)^{\lambda_1 - 1}(1 - x)^{\lambda_1 + \lambda_2 - 2} = \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, x^{\lambda_1 - 1}(1 - x)^{\lambda_2 - 1}, \quad x \in [0, 1].$$

Pour les moments de β, on a

$$\mathbf{E}\beta^l = \frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)} \int_0^1 x^{\lambda_1 + l - 1}(1 - x)^{\lambda_2 - 1} dx = \frac{\Gamma(\lambda_1 + \lambda_2)\Gamma(\lambda_1 + l)}{\Gamma(\lambda_1)\Gamma(\lambda_1 + \lambda_2 + l)}.$$

Pour $l = 1, 2$, il vient

$$\mathbf{E}\beta = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \quad \mathbf{E}\beta^2 = \frac{\lambda_1(\lambda_1 + 1)}{(\lambda_1 + \lambda_2)(\lambda_1 + \lambda_2 + 1)}.$$

9. Distribution uniforme. La distribution uniforme sur $[0, 1]$ est un cas particulier de la distribution bêta pour $\lambda_1 = \lambda_2 = 1$.

On désignera $\mathbf{U}_{a,b}$ la distribution uniforme sur $[a, b]$, de sorte que $\mathbf{B}_{1,1} = \mathbf{U}_{0,1}$.

La distribution bêta permet de décrire la loi des termes $x_{(k)}$ de l'échantillon ordonné associé à un échantillon X.

THÉORÈME 1. *Si X est un échantillon de distribution* $\mathbf{P}$ *et de fonction de répartition F continue, on a*

$$y_{(k)} = F(x_{(k)}) \in \mathbf{B}_{k,\, n-k+1}.$$

DÉMONSTRATION. Puisque $y_k = F(x_k) \in \mathbf{U}_{0,1}$, on peut traiter $y_{(k)} = F(x_{(k)})$ comme un terme de l'échantillon ordonné associé à un échantillon $Y \in \mathbf{U}_{0,1}$. Trouvons $\mathbf{P}(y_{(k)} \in]u, u + du[)$. L'événement $\{y_{(k)} \in]u, u + du[\}$ peut être représenté comme la réunion des événements disjoints

$$A_j = \{y_j \in]u, u + du[,\ y_j = y_{(k)}\},$$

qui se produisent lorsque y_j tombe dans $]u, u + du[$ (avec la probabilité du), $k - 1$ des $n - 1$ observations restantes tombent dans l'intervalle $]0, u[$ et $n - k$ observations dans l'intervalle $]u, 1[$. Donc

$$\mathbf{P}(A_j) = C_{n-1}^{k-1}\, u^{k-1}(1 - u)^{n-k} du,$$

$$\mathbf{P}(y_{(k)} \in]u, u + du[) = nC_{n-1}^{k-1}\, u^{k-1}(1 - u)^{n-k} du.$$

Ceci exprime que la densité de $y_{(k)}$ existe et vaut

$$\frac{n!}{(k-1)!(n-k)!}\, u^{k-1}(1-u)^{n-k} = \frac{\Gamma(n + 1)}{\Gamma(k)\Gamma(n-k+1)}\, u^{k-1}(1-u)^{n-k}. \blacktriangleleft$$

En se servant du théorème 1, on peut obtenir sans peine la distribution *limite* des termes de l'échantillon ordonné lorsque la taille de X croît indéfiniment. Nous nous arrêterons sur un seul résultat découlant des théorèmes de continuité.

THÉORÈME 2. *Si* $a = \dfrac{k}{n + 1} \to a_0 \in]0, 1[$ *pour* $n \to \infty$, *on a*

$$y_{(k)} = a + \frac{\sqrt{a_0(1 - a_0)}}{\sqrt{n}}\xi_n, \quad \xi_n \in \Phi_{0,1}.$$

DÉMONSTRATION. Le théorème 1 nous dit que $y_{(k)} \in B_{k,\, n-k+1}$, donc, en vertu des propriétés de la distribution bêta, on a la représentation

$$y_{(k)} \underset{d}{=} \frac{\eta_1}{\eta_1 + \eta_2}, \quad \eta_j \in H_{k_j}, \quad k_1 = 2k,\ k_2 = 2(n - k + 1).$$

Posons par souci de simplicité $a_1 = a$, $a_2 = 1 - a$ et supposons que $a = a_0$ fixe. Il est alors évident que $k_j/(n + 1) = 2a_j$, $j = 1, 2$, et en vertu des propriétés de la distribution χ^2

$$\eta_j = k_j + \sqrt{2k_j}\, \xi_n^{(j)}, \quad \xi_n^{(j)} \Rightarrow \xi^{(j)} \in \Phi_{0,\,1};$$

$$y_{(k)} = \frac{a_1 + \sqrt{\dfrac{a_1}{n+1}}\, \xi_n^{(1)}}{a_1 + a_2 + \sqrt{\dfrac{a_1}{n+1}}\, \xi_n^{(1)} + \sqrt{\dfrac{a_2}{n+1}}\, \xi_n^{(2)}}.$$

Reste à appliquer le théorème de continuité 1.5.3A pour

$$H(t) = \frac{t_1}{t_1 + t_2}, \quad b_n = \frac{1}{\sqrt{n+1}}, \quad \eta_n^{(j)} = \sqrt{a_j}\, \xi_n^{(j)}.$$

Comme η_j (donc $\xi_n^{(j)}$) sont indépendantes et

$$\frac{\partial H}{\partial t_1} = \frac{t_2}{(t_1 + t_2)^2}, \quad \frac{\partial H}{\partial t_2} = -\frac{t_1}{(t_1 + t_2)^2},$$

il vient

$$(y_{(k)} - a_1)\sqrt{n+1} \Rightarrow a_2\sqrt{a_1}\, \xi^{(1)} - a_1\sqrt{a_2}\, \xi^{(2)} \underset{d}{=} \sqrt{a_1 a_2}\, \xi, \ \xi \in \Phi_{0,\,1}.$$

Si a dépend de n, il faut se servir de la remarque 1.5.1. ◄

COROLLAIRE 1. *Si $a = k/(n + 1) \to a_0 \in]0, 1[$ et la fonction continue F est continûment dérivable au point $\zeta_0 = F^{-1}(a_0)$ (le quantile d'ordre a_0), alors*

$$x_{(k)} = \zeta + \frac{\sqrt{a_0(1 - a_0)}\, \xi_n}{f(\zeta_0)\sqrt{n}}, \quad \xi_n \in \Phi_{0,\,1}, \tag{11}$$

où $\zeta = F^{-1}(a)$ est le quantile d'ordre a, $f(x) = F'(x)$.

Cette proposition dérive directement du théorème de continuité 1.5.3 (compte tenu de la remarque 1.5.1) si l'on se sert de la représentation

$$x_{(k)} = F^{-1}(y_{(k)}) = F^{-1}\left(a + \sqrt{\frac{a_0(1 - a_0)}{n}}\,\xi_n\right)$$

et du fait que $\dfrac{dF^{-1}(x)}{dx} = \dfrac{1}{f(F^{-1}(x))}$.

REMARQUE 1. La proposition (11) généralise un peu le corollaire 1.8.1. On peut aussi le généraliser dans une autre direction. Supposons que pour $x \to \zeta$

$$|F(x) - F(\zeta)| \sim c|x - \zeta|^{\gamma}, \quad \gamma > 0.$$

Il est immédiat de voir que pour $y \to a$

$$|F^{-1}(y) - F^{-1}(a)| \sim \left|\frac{y - a}{c}\right|^{1/\gamma},$$

donc

$$(x_{(k)} - \zeta)n^{\frac{1}{2\gamma}} \Rightarrow (a_0(1 - a_0))^{\frac{1}{2\gamma}} |\xi/c|^{\frac{1}{\gamma}} \operatorname{sgn}\xi, \quad \xi \in \Phi_{0,1}. \qquad (12)$$

Ce qui entraîne (11) pour $\gamma = 1$ et $c = f(\zeta)$.

10. Distribution $\mathbf{K}_{\alpha,\sigma}$ de Cauchy de paramètres (α, σ). On appelle ainsi la distribution de densité

$$k_{\alpha,\sigma}(x) = \frac{\sigma}{\pi[\sigma^2 + (x - \alpha)^2]} = \frac{1}{\pi\sigma} \cdot \frac{1}{1 + \left(\dfrac{x - \alpha}{\sigma}\right)^2}.$$

Comme pour la loi normale, les paramètres α et σ sont ici respectivement les paramètres de localisation et d'échelle. La forme de la distribution $\mathbf{K}_{0,1}$ rappelle beaucoup celle de $\Phi_{0,1}$, mais la densité $k_{0,1}$ présente, comme la densité de la loi de Student, des ailes plus « larges » (c'est-à-dire décroît plus lentement pour $|x| \to \infty$), de sorte que la distribution $\mathbf{K}_{0,1}$ ne possède même pas d'espérance mathématique finie. Dans [11], chap. 7 on a signalé que les distributions $\mathbf{K}_{\alpha,\sigma}$ sont stables comme les distributions normales. La fonction caractéristique $\varkappa_{0,1}(t)$ de la distribution $\mathbf{K}_{0,1}$ est

$$\varkappa_{0,1}(t) = e^{-|t|},$$

donc

$$\varkappa_{\alpha,\sigma}(t) = \exp\{i\alpha t - \sigma|t|\},$$

$$\varkappa_{\alpha_1,\sigma_1}(t)\varkappa_{\alpha_2,\sigma_2}(t) = \exp\{i(\alpha_1 + \alpha_2)t - (\sigma_1 + \sigma_2)|t|\},$$

de sorte que le produit de convolution de $\mathbf{K}_{\alpha_1,\sigma_1}$ et de $\mathbf{K}_{\alpha_2,\sigma_2}$ est égal à $\mathbf{K}_{\alpha_1+\alpha_2,\sigma_1+\sigma_2}$.

Il est immédiat de remarquer que $\mathbf{K}_{0,1} = \mathbf{T}_1$.

Dans les applications, on a souvent affaire aux fonctions de variables aléatoires normales. L'une d'elles est la fonction exponentielle à laquelle est reliée la distribution log-normale.

11. Distribution log-normale $\mathbf{L}_{\alpha,\sigma^2}$. On dira que $\eta \in \mathbf{L}_{\alpha,\sigma^2}$ si $\ln\eta \in \in \Phi_{\alpha,\sigma^2}$. Autrement dit, $\eta = e^{\xi}$, où $\xi \in \Phi_{\alpha,\sigma^2}$. On voit que la distribution $\mathbf{L}_{\alpha,\sigma^2}$ est concentrée sur le demi-axe positif.

En vertu des formules de la densité d'une fonction d'une variable aléatoire (cf. [11]), la densité de $\eta \in \mathbf{L}_{\alpha,\sigma^2}$ est égale à

$$\varphi_{\alpha,\sigma^2}(\ln x)\, x^{-1}.$$

Par ailleurs,

$$\mathbf{E}\eta = \int e^y \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\alpha)^2}{2\sigma^2}} dy =$$

$$= \exp \frac{(\alpha+\sigma^2)^2 - \alpha^2}{2\sigma^2} \int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\alpha-\sigma^2)^2}{2\sigma^2}\right) dy = e^{\alpha+\sigma^2/2},$$

$$\mathbf{E}\eta^2 = \int e^{2y} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\alpha)^2}{2\sigma^2}} dy = e^{2\alpha+2\sigma^2}.$$

12. Distribution dégénérée. Le symbole $\mathbf{I}_a$ (que nous avons déjà utilisé dans le § 1.2) désignera une distribution dégénérée concentrée en un point a.

Dans le cas général où l'on étudiera une *famille arbitraire de distributions* dépendant d'un paramètre θ (scalaire ou vectoriel), on se servira de la notation $\mathbf{P}_\theta$. La famille, quant à elle, sera désignée par

$$\{\mathbf{P}_\theta\}_{\theta\in\Theta},$$

où Θ est l'ensemble de toutes les valeurs possibles du paramètre θ. On appliquera les mêmes notations aux familles de distributions 1 à 12. Ainsi, $\{\Phi_{\alpha,1}\}_{\alpha\in R}$ désignera la famille de toutes les distributions normales de variance 1.

Les distributions 1 à 11 sont absolument continues par rapport à la mesure de Lebesgue. Introduisons maintenant les notations de trois distributions *discrètes* bien connues (absolument continues par rapport à la mesure cardinale $\mu(B)$ définie comme suit : $\mu(B) = k$ si B contient k points entiers).

13. Distribution $\mathbf{B}_p^n$ de Bernoulli. Par définition, $\xi \in \mathbf{B}_p^n$ (n est entier, $p \in [0, 1]$) si

$$\mathbf{P}(\xi = k) = C_n^k p^k (1-p)^{n-k}, \quad 0 \leqslant k \leqslant n.$$

14. Distribution Π_λ de Poisson. Cette distribution est définie par la relation

$$\Pi_\lambda(B) = \sum_{\substack{k \in B \\ k \geq 0}} \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0.$$

15. Distribution polynomiale. Nous la désignerons par $\mathbf{B}_p^n$, où $n > 0$ est un entier, $p = (p_1, \ldots, p_r)$, $p_j \geqslant 0$, $\sum_{j=1}^{r} p_j = 1$. Etant donné un vecteur aléatoire à composantes entières $\nu = (\nu_1, \ldots, \nu_r)$, on écrira $\nu \in \mathbf{B}_p^n$ si pour $k = (k_1, \ldots, k_r)$, $k_j \geqslant 0$, $\sum_{j=1}^{r} k_j = n$, on a

$$\mathbf{P}(\nu = k) = \frac{n!}{k_1! \ldots k_r!} p_1^{k_1} \ldots p_r^{k_r}.$$

La distribution $\mathbf{B}_p^n$ est rattachée à une suite de n épreuves indépendantes donnant lieu à une issue sur r possibles incompatibles $A_1, \ldots, A_r$, la probabilité d'apparition de l'issue A_j au cours d'une épreuve étant égale à p_j. Les coordonnées ν_j du vecteur ν représentent les fréquences d'apparition des événements A_j au terme de n épreuves (cf. par exemple [11]). Il est évident que pour chaque $j = 1, \ldots, r$

$$\nu_j \in \mathbf{B}_{p_j}^n.$$

L'issue de la j-ième épreuve peut être décrite par un vecteur x_j à r dimensions dont une composante est égale à 1 et les $r - 1$ autres à 0. Le numéro de cette composante est le numéro de l'événement qui s'est produit au cours de la j-ième épreuve. Il est évident que $\nu = \sum_{j=1}^{n} \mathrm{x}_j$. S'agissant d'un échantillon $X = (\mathrm{x}_1, \ldots, \mathrm{x}_n)$, il nous sera plus commode de noter

$$X \in \mathbf{B}_p,$$

où $\mathbf{B}_p = \mathbf{B}_p^1$. L'espace $\mathscr{X}$ attaché à cet échantillon est visiblement fini et composé de r points. Si $p = (p_1, p_2)$, $p_1 + p_2 = 1$, on obtient le schéma de Bernoulli pour lequel nous emploierons les mêmes notations en identifiant $\mathbf{B}_{(p_1, p_2)}$ à $\mathbf{B}_{p_1} = \mathbf{B}_{p_1}^1$ (cf. n° 13). Dans le cas général, la distribution $\mathbf{B}_p$ ne dépend en fait que d'un paramètre $(p_1, \ldots, p_{r-1})$, de sorte qu'on aurait pu remplacer l'indice p par $(p_1, \ldots, p_{r-1})$.

Parmi les distributions envisagées ci-dessus, plusieurs, notamment $\Phi_{0,1}$, $\mathbf{H}_k$, $\mathbf{F}_{k_1, k_2}$, $\mathbf{T}_k$, Π_λ, sont tabulées dans des aide-mémoire de statistique mathématique et dans des tables spéciales (cf. par exemple [8]).

§ 3. Estimation ponctuelle. Méthode fondamentale d'estimation. Convergence, normalité asymptotique

1. Méthode de substitution. Convergence. La notion d'estimateur a été introduite au § 1. Formellement c'est la même chose qu'une statistique, c'est-à-dire une fonction mesurable θ^* d'un échantillon. De façon non formelle, nous appelons *estimateurs* θ^* les seules statistiques qui sont destinées à remplacer le paramètre inconnu θ. En d'autres termes, θ^* est une approximation de θ dépendant de l'échantillon. Une valeur de θ^* est appelée estimation *ponctuelle* de θ par opposition aux estimations *par intervalles* qui seront envisagées plus bas.

La donnée d'un estimateur suppose généralement la donnée de fonctions (des échantillons X_n) définies pour toutes les valeurs possibles de n. Aussi dans la suite, le terme « estimateur » désignera-t-il une famille de statistiques $\theta^* = \theta_n^*(X_n)$ définies pour tous les $n = 1, 2, \ldots$, où θ^* est une fonction sur $\mathscr{X}^n$, ou, ce qui revient au même, une fonction $\theta^* = \theta^*(n, X_\infty)$ définie sur le produit de l'ensemble des entiers par $\mathscr{X}^\infty$.

Conformément au § 1, nous admettrons que dans la position du problème d'estimation sont définis l'ensemble Θ des valeurs possibles du paramètre θ et la famille $\mathscr{P}$ des distributions possibles $\mathbf{P}$ de l'échantillon X (ce peuvent être, disons, seulement les distributions normales $\Phi_{\alpha,1}$ ou les distributions de Poisson Π_λ dont on demande d'estimer les paramètres inconnus α, λ). Si aucune condition n'est imposée à θ (ou à $\mathbf{P}$), on pourra admettre que Θ (resp. $\mathscr{P}$) est confondu avec l'espace euclidien de dimension correspondante (resp. avec l'ensemble de toutes les distributions).

Pour désigner l'estimateur d'un paramètre, on fera suivre le symbole de ce dernier par un astérisque. Par exemple, un estimateur du paramètre α d'une loi normale sera

$$\alpha^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Les moments empiriques utilisés pour estimer

$$\mathbf{E}\,x_1 = \int x\mathbf{P}(dx) \quad \text{et} \quad \mathbf{V}x_1 = \int (x - \mathbf{E}x_1)^2\mathbf{P}(dx)$$

sont désignés par des notations spéciales :

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Nous avons déjà signalé qu'un paramètre possédait une foule d'estimateurs et avant de discuter les critères de leur qualité dans telle ou telle situa-

tion, arrêtons-nous sur quelques méthodes générales « régulières » de leur construction.

Ces méthodes regroupent les approches les plus logiques du problème d'estimation et nous permettent d'acquérir les meilleurs estimateurs selon tel ou tel critère.

Les méthodes d'estimation sont presque toutes basées sur la *substitution d'une distribution empirique.*

Soit $X_n \in \mathbf{P}$ et supposons que le paramètre inconnu θ se représente par une fonctionnelle G de la distribution $\mathbf{P}$:

$$\theta = G(\mathbf{P}).$$

Supposons par ailleurs que $\mathbf{P}_n^*$ est comme toujours une distribution empirique. La méthode de substitution nous commande de prendre pour estimateur θ^* la fonction

$$\theta^* = G(\mathbf{P}_n^*).$$

Un tel estimateur sera appelé *estimateur par la méthode de substitution* ou pour simplifier *estimateur de substitution.*

La fonctionnelle G est parfois donnée sous forme implicite comme solution d'une équation $H(\theta, \mathbf{P}) = 0$ résoluble en θ. Dans ce cas, aux termes de la définition fondamentale, on appelle estimateur de substitution toute solution de l'équation $H(\theta, \mathbf{P}_n^*) = 0$.

Si l'on sait que le paramètre $\theta \in R^k$ prend ses valeurs dans un domaine $\Theta \subset R^k$, on peut utiliser cette information dans la construction des estimateurs de substitution. Supposons que le domaine Θ est fermé et soit $\mathscr{P}$ l'ensemble de toutes les distributions de l'échantillon X, $\Theta = \{G(\mathbf{P})\}_{\mathbf{P}\in\mathscr{P}}$. Définissons la fonctionnelle $G_1(\mathbf{P})$, où $\mathbf{P}$ est arbitraire, comme la valeur $t \in \Theta$ pour laquelle

$$\min_{t\in\Theta} |t - G(\mathbf{P})| = |G_1(\mathbf{P}) - G(\mathbf{P})|, \qquad (1)$$

de sorte que $G_1(\mathbf{P})$ est le point de Θ le plus proche de $G(\mathbf{P})$. Puisque $G_1(\mathbf{P}) = G(\mathbf{P}) = \theta$ si $\mathbf{P} \in \mathscr{P}$, l'estimateur

$$\theta^* = G_1(\mathbf{P}_n^*) \qquad (2)$$

sera avec $G(\mathbf{P}_n^*)$ un estimateur de substitution et de plus l'ensemble des valeurs possibles de θ^* sera inclus dans Θ.

Au sujet des estimateurs (1) et (2) on dira qu'ils ont été acquis par *restriction de la méthode de substitution.*

Supposons maintenant qu'on estime le paramètre α de la loi normale $\Phi_{\alpha,1}$ et que l'on sache *a priori* que $\alpha \in [0, 1]$. Il est alors possible que $\alpha^* = \bar{x} \notin [0, 1]$ (il est évident que $\bar{x} = \int t\, dF_n^*(t)$ est une estimation de substitu-

tion). La restriction de la méthode de substitution nous suggère de prendre pour estimation le point de [0, 1] le plus proche de $\bar{x}$.

Signalons maintenant que telle qu'elle a été formulée la méthode de substitution n'a pas toujours de sens. En effet, la fonctionnelle G peut ne pas être définie sur l'ensemble des distributions empiriques. Supposons par exemple que l'on sache *a priori* que la distribution $\mathbf{P}$ appartient à la classe $\mathscr{P}$ des distributions absolument continues par rapport à la mesure de Lebesgue, de sorte que chaque $\mathbf{P} \in \mathscr{P}$ admet une densité f. On s'intéresse à la valeur

$$\theta = G(\mathbf{P}) = \int f^2(x)dx = \int \left(\frac{d\mathbf{P}}{dx}\right)^2 dx.$$

Il est clair que $G(\mathbf{P}_n^*)$ n'a pas de sens dans ce cas, puisque $\mathbf{P}_n^*$ est une distribution discrète. Dans ces cas la méthode de substitution peut toujours être modifiée naturellement de manière à garder son sens. Dans l'exemple cité où $G(\mathbf{P})$ est une fonctionnelle de la densité f, pour θ^* il faut envisager, en vertu de la méthode de substitution, la valeur $G(\mathbf{P}_n^{**})$, où $\mathbf{P}_n^{**}$ est une distribution empirique lissée (cf. § 1.10) dont la densité converge vers $f(x)$.

Il est possible aussi que parfois $G(\mathbf{P}_n^*)$ n'ait pas de sens pour tous les X_n, mais seulement pour $X_n \in A_n$, où $\mathbf{P}(X_n \in A_n) \to 1$ pour $n \to \infty$. Cette circonstance est sans conséquence sur la suite et pour fixer les idées on peut poser $G(\mathbf{P}_n^*) = 0$ pour $X_n \notin A_n$.

Dans ce paragraphe on admettra pour simplifier que $G(\mathbf{P}_n^*)$ a un sens pour tous les $X_n \in \mathscr{X}^n$ et que θ^* est une variable aléatoire, c'est-à-dire que $G(\mathbf{P}_n^*)$ est une application mesurable de $\mathscr{X}^n$ dans R^k, où k est la dimension de θ.

Le principe de substitution est une approche assez naturelle du problème, puisque, comme on sait déjà, la distribution $\mathbf{P}_n^* \to \mathbf{P}$ lorsque $n \to \infty$.

Supposons que $X_n = [X_\infty]_n$.

DÉFINITION 1. On dit qu'un estimateur $\theta^* = \theta_n^*(X_n)$ (ou une suite $\theta_n^*(X)$) est *convergent* si

$$\theta^* \underset{P}{\to} \theta$$

lorsque $n \to \infty$.

Un estimateur θ^* est *fortement convergent* si

$$\theta^* \underset{\text{p.s.}}{\to} \theta$$

lorsque $n \to \infty$.

Supposons comme toujours que F est la fonction de répartition de $\mathbf{P}$.

THÉORÈME 1. *Supposons que* $\theta = G(\mathbf{P})$ *et que la fonctionnelle* G *appartienne à l'une des deux classes suivantes : soit elle se représente sous la forme*

$$G(\mathbf{P}) = h\left(\int g(x)dF(x)\right), \tag{I}$$

où h est une fonction continue au point $a = \int g(x)dF_0(x)$ *(et c'est une fonctionnelle de type* I), *soit elle se représente sous la forme*

$$G(\mathbf{P}) = G_1(F), \tag{II}$$

où la fonctionnelle G_1 *est continue au point* F_0 *par rapport à la métrique uniforme (et c'est une fonctionnelle de type* II). *Dans ces conditions, si* $X \in F_0$, *alors* $\theta^* = G(\mathbf{P}_n^*)$ *est un estimateur fortement convergent :*

$$\theta^* \underset{\text{p.s.}}{\rightarrow} \theta.$$

Ce théorème résulte immédiatement du théorème 1.4.1.

2. Normalité asymptotique. Cas d'un paramètre scalaire.

DÉFINITION 2. On dit qu'un estimateur θ^* d'un paramètre θ est *asymptotiquement normal de paramètre* $\sigma^2 \geqslant 0$ si $(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\,\sigma^2}$.

La dernière relation peut être lue de la manière suivante : l'estimateur θ^* est asymptotiquement normal de paramètres $(\theta, \sigma^2/n)$.

Supposons que θ^* est un estimateur de substitution du paramètre $\theta = G(\mathbf{P})$ et que (I) est réalisée, c'est-à-dire que

$$\theta^* = h\left(\frac{1}{n}\sum_{i=1}^{n} g(x_i)\right) \tag{3}$$

est une statistique du premier type. Les résultats du § 1.7 entraînent la proposition suivante. Supposons que θ est un paramètre scalaire et g, une fonction scalaire.

THÉORÈME 2. *Si* $X \in F_0$, *h est dérivable au point* $a = \int g(x)dF_0(x)$, $0 < |h'(a)| < \infty$, $\int g^2(x)dF_0(x) < \infty$, *alors* θ^* *est un estimateur asymptotiquement normal de paramètre*

$$\sigma^2 = [h'(a)]^2 \int (g(x) - a)^2 dF_0(x).$$

Les exemples traités dans le § 1.7 peuvent servir à illustrer ce théorème, puisque les statistiques mises en jeu sont utilisées pour estimateurs.

Par analogie, on aurait pu se servir des résultats du § 1.8 pour établir les conditions de normalité asymptotique d'estimateurs qui sont des statistiques du deuxième type. Le lecteur peut déduire les assertions nécessaires à l'aide du théorème 1.8.1 en y exigeant seulement que $k = 1$ et que la dérivée g soit telle que $g(F_0, w^\circ) \in \Phi_{0,\,\sigma^2}$.

3. Normalité asymptotique. Cas d'un paramètre vectoriel.

DÉFINITION 2A. On dit qu'un estimateur $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ de $\theta = (\theta_1, \ldots, \theta_k)$ est un *estimateur asymptotiquement normal de matrice* σ^2 si

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\,\sigma^2}, \tag{4}$$

où Φ_{0,σ^2} est une distribution normale k-dimensionnelle d'espérance mathématique nulle et de matrice des moments d'ordre deux $\sigma^2 = \|\sigma_{ij}\|$. La densité de cette distribution est (cf. § 2)

$$\varphi_{0,\sigma^2}(x) = \frac{\sqrt{|A|}}{(2\pi)^{k/2}} e^{-\frac{1}{2}xAx^T},$$

où A est la matrice inverse de σ^2, $x = (x_1, \ldots, x_k)$.

Si θ^* est un estimateur de substitution et s'il est unė statistique du premier type (c'est-à-dire représentable sous la forme (3), où g, θ^* et h sont des fonctions vectorielles), on peut se servir du théorème 1.7.1A et de la remarque qui le suit pour établir les conditions de normalité asymptotique.

THÉORÈME 2A. *Supposons que $\theta^* \in R^k$ est défini par* (I), *où $g = (g_1, \ldots, g_s) \in R^s$, et que la fonction vectorielle $h(t) = (h_1(t), \ldots, h_k(t))$, $t = (t_1, \ldots, t_s)$, admet au point $a = (a_1, \ldots, a_s)$, $a_j = \int g_j(x)dF_0(x)$, les dérivées partielles $\frac{\partial h_l}{\partial t_j}(a)$, $l = 1, \ldots, k$, $j = 1, \ldots, s$. Sous ces conditions, si $X \in F_0$, alors*

$$(\theta^* - \theta)\sqrt{n} \Rightarrow \xi H^T,$$

où $\xi = (\xi_1, \ldots, \xi_s) \in \Phi_{0,d^2}$ est un vecteur normal de moyenne nulle et de matrice des moments d'ordre deux $d^2 = \|d_{ij}\|$, $d_{ij} = \mathbf{E}(g_i(x_1) - a_i)(g_j(x_1) - a_j)$, $i, j = 1, \ldots, s$; $H = \|h_{ij}\|$ est une (k, s)-matrice d'éléments $h_{ij} = \frac{\partial h_i}{\partial t_j}(a)$, $i = 1, \ldots, k$; $j = 1, \ldots, s$.

Ceci exprime que si les conditions du théorème 2A sont remplies, l'estimateur θ^* est un estimateur asymptotiquement normal de matrice $\sigma^2 = Hd^2H^T = \mathbf{E}H\xi^T\xi H^T$. Signalons que les dimensions k et s des matrices σ^2 et d^2 sont différentes ici.

§ 4. Réalisation de la méthode de substitution dans le cas paramétrique. Méthode des moments

Supposons que $X \in \mathbf{P}_\theta$, où $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ est une famille donnée de distributions $\mathbf{P}_\theta$ dépendant d'un paramètre θ. Le paramètre θ peut être aussi bien scalaire que vectoriel. Si par exemple $X \in \Phi_{\alpha,\sigma^2}$, alors $\theta = (\alpha, \sigma^2)$ est à deux dimensions et l'ensemble Θ est le demi-plan $\{-\infty < \alpha < \infty, \sigma \geqslant 0\}$ ou l'une quelconque de ses parties.

L'espérance mathématique et la variance d'une statistique $S = S(X)$, où $X \in \mathbf{P}_\theta$, seront désignées respectivement par $\mathbf{E}_\theta S$ et $\mathbf{V}_\theta S$.

On se propose d'étudier quelques méthodes d'estimation dont chacune

peut être traitée comme une réalisation du principe de substitution d'une distribution empirique.

1. Méthode des moments. Cas scalaire. Choisissons $g(x)$ de telle sorte que la fonction

$$m(\theta) = \mathbf{E}_\theta g(x_1) = \int g(x)\mathbf{P}_\theta(dx) \quad (1)$$

soit monotone et continue. Le domaine $m(\Theta)$ des valeurs de $m(\theta)$, $\theta \in \Theta$, sera de même « nature » que Θ. Si par exemple Θ est un intervalle de l'axe réel, il en sera de même de $m(\Theta)$.

Il est évident que l'équation $m(\theta) = t$ admet une seule solution continue dans le domaine $m(\Theta)$, soit $\theta = m^{-1}(t)$, et que (1) peut être écrit sous la forme équivalente :

$$\theta = m^{-1}\left(\int g(x)\mathbf{P}_\theta(dx)\right). \quad (2)$$

Supposons pour simplifier que

$$\bar{g} \equiv \int g(x)d\mathbf{P}_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i) \in m(\Theta)$$

pour tous les $X \in \mathscr{X}^n$.

DÉFINITION 1. On appelle *estimateur par la méthode des moments l'estimateur*

$$\theta^* = m^{-1}(\bar{g}). \quad (3)$$

Si $\bar{g} \notin m(\Theta)$, on peut, en vertu de (3.1) et (3.2), poser

$$\theta^* = m^{-1}(\bar{g}_0),$$

où $\bar{g}_0 \in m(\Theta)$ est le point de $m(\Theta)$ le plus proche de $\bar{g}$.

Il est aisé de voir que c'est un estimateur de substitution. Le choix de la fonction $m(\theta)$ nous a permis d'exprimer θ sous la forme d'une fonctionnelle (2). Il est clair aussi que l'estimateur (3) est une statistique du premier type, de sorte que les estimateurs par la méthode des moments seront fortement convergents en vertu du théorème 3.1. Si en outre la fonction m est dérivable au point θ et $\int g^2(x)\mathbf{P}_\theta(dx) < \infty$, le théorème 3.2 nous dit que l'estimateur par la méthode des moments sera asymptotiquement normal de paramètre $(m'(\theta))^{-2}\mathbf{V}_\theta g(x_1)$.

La méthode des moments a été proposée par Pearson (dans une forme plus particulière) et historiquement est la première méthode régulière de construction des estimateurs.

Cette méthode tient son nom du fait qu'elle consiste à égaler les moments « théoriques » et empiriques (les espérances mathématiques) de la variable $g(x_1)$: en effet, l'estimateur (3) n'est autre que la solution de

l'équation

$$m(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(x_i). \tag{4}$$

Ajoutons que pour $g(x)$ on prend souvent la fonction $g(x) = x$ ou $g(x) = x^k$, $k > 1$, de sorte que notre équation se transforme en équation pour les moments ordinaires.

L'égalité (4) peut également être traitée comme une égalité entre les moyennes de $g(x_1)$ sur l' « espace » et sur le « temps ».

La non-univocité de la méthode des moments, de même que du principe de substitution, saute aux yeux : en effet, le choix de la fonction $g(x)$ n'a été soumis pratiquement à aucune contrainte.

EXEMPLE 1. Soit $X \in \Gamma_{\alpha,\,1}$ et supposons que α est inconnu. Construisons des estimateurs par la méthode des moments avec les deux fonctions élémentaires $g_1(x) = x$ et $g_2(x) = x^2$. On a les égalités suivantes (cf. n°5, § 2)

$$m_1(\alpha) = \mathbf{E}_\alpha g_1(x_1) = \int_0^\infty x\Gamma_{\alpha,\,1}(dx) = 1/\alpha,$$

$$m_2(\alpha) = \mathbf{E}_\alpha g_2(x_1) = \int_0^\infty x^2\Gamma_{\alpha,\,1}(dx) = 2/\alpha^2.$$

La résolution des équations $m_1(\alpha) = \bar{x}$, $m_2(\alpha) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, nous donne les estimateurs par la méthode des moments

$$\alpha^* = (\bar{x})^{-1}, \quad \alpha^{**} = \left(\frac{1}{2n}\sum_{i=1}^{n} x_i^2\right)^{-1/2} \tag{5}$$

Ces deux estimateurs sont des statistiques du premier type et nous pouvons décrire leurs propriétés asymptotiques. On obtient en vertu des égalités (2.4) :

$$\mathbf{V}_\alpha g_1(x_1) = \mathbf{V}_\alpha x_1 = 1/\alpha^2, \quad \mathbf{V}_\alpha g_2(x_1) = \mathbf{V}_\alpha x_1^2 = 20/\alpha^4.$$

Comme $m_1'(\alpha) = -1/\alpha^2$ pour le premier estimateur et $m_2'(\alpha) = -4/\alpha^3$ pour le second, les théorèmes 3.1 et 3.2 nous disent que les estimateurs α^* et α^{**} sont fortement convergents et asymptotiquement normaux de paramètres

$$\frac{1}{\alpha^2}\cdot\alpha^4 = \alpha^2, \quad \frac{20}{\alpha^4}\cdot\frac{\alpha^6}{16} = \frac{5}{4}\alpha^2$$

respectivement.

L'estimateur α^* est visiblement meilleur, puisque la variance de sa distribution limite est plus petite.

2. Méthode des moments. Cas vectoriel. Le cas où θ est un paramètre vectoriel se traite exactement de la même manière.

Supposons comme toujours que θ est un paramètre à k dimensions. Choisissons la fonction vectorielle $g(x) = (g_1(x), \ldots, g_k(x))$ de telle sorte que l'équation

$$m(\theta) = t,$$

où

$$t = (t_1, \ldots, t_k), \quad m(\theta) = (m_1(\theta), \ldots, m_k(\theta)),$$
$$m_j(\theta) = \mathbf{E}_\theta g_j(x_1) = \int g_j(x) \mathbf{P}_\theta(dx),$$

admette une solution continue unique $\theta = m^{-1}(t)$ dans le domaine $m(\Theta)$ des valeurs de $m(\theta)$, $\theta \in \Theta$. Supposons pour simplifier que le vecteur

$$\bar{g} = \left(\frac{1}{n}\sum_{i=1}^{n} g_1(x_i), \ldots, \frac{1}{n}\sum_{i=1}^{n} g_k(x_i)\right)$$

appartient au domaine $m(\Theta)$ pour tous les $X \in \mathscr{X}^n$.

DÉFINITION 1A. L'estimateur $\theta^* = m^{-1}(\bar{g})$ s'appelle *estimateur par la méthode des moments.*

Comme précédemment le théorème 3.1 nous dit que ces estimateurs seront fortement convergents.

Pour que l'estimateur θ^* soit asymptotiquement normal, il faut exiger en outre que la fonction m soit dérivable, $\int g_j^2(x)\mathbf{P}_\theta(dx) < \infty$. Le théorème 3.2A nous permet d'obtenir sans peine une assertion sur la distribution limite de θ^*.

EXEMPLE 2. Prenons pour $\{\mathbf{P}_\theta\}$ la famille des distributions normales Φ_{α, σ^2}. En admettant que $g_1(x) = x$ et $g_2(x) = x^2$, on obtient les équations suivantes de la méthode des moments

$$\alpha = \bar{x}, \quad \sigma^2 + \alpha^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

dont les solutions sont

$$\alpha^* = \bar{x}, \quad (\sigma^2)^* = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = S^2.$$

Nous proposons au lecteur de trouver, à titre d'exercice, les estimateurs, par la méthode des moments, des paramètres pour toutes les familles paramétriques citées dans le § 2.

3. Méthode des moments généralisée. La méthode des moments admet la généralisation suivante qui élargit considérablement la classe des estimateurs envisagée ci-dessus. Bornons-nous, pour simplifier, au cas d'un paramètre θ scalaire. Considérons une fonction de deux variables $g(x, \theta)$ et supposons que pour toute distribution $\mathbf{P}$ l'équation

$$\int g(x, \theta)\mathbf{P}(dx) = \int g(x, \theta)\mathbf{P}_\theta(dx) \tag{6}$$

admet une solution $\theta = G(\mathbf{P})$, de sorte que $\theta = G(\mathbf{P}_\theta)$ pour $\mathbf{P} = \mathbf{P}_\theta$.

On appellera *estimateur par la méthode des moments généralisée* l'estimateur

$$\theta^* = G(\mathbf{P}_n^*).$$

Il est évident que c'est un estimateur de substitution comme les estimateurs par la méthode des moments. L'étude des propriétés de ces estimateurs est plus compliquée. Nous aurons l'occasion de nous en assurer dans les prochains paragraphes dans la mesure où l'un des estimateurs de substitution que nous aurons à étudier en détail sera un estimateur par la méthode des moments généralisée.

§ 5*. Méthode de la distance minimale

Cette méthode qui, comme celle des moments, est une réalisation du principe de substitution consiste en ce qui suit. Considérons une fonctionnelle de deux distributions $d(\mathbf{P}, \mathbf{Q})$ qui, regardée comme une fonction de $\mathbf{Q}$, atteint son minimum pour $\mathbf{Q} = \mathbf{P}$ et $d(\mathbf{P}, \mathbf{Q}) > d(\mathbf{P}, \mathbf{P})$ pour $\mathbf{Q} \neq \mathbf{P}$. Nous traiterons la quantité $d(\mathbf{P}, \mathbf{Q})$ (ou $d(\mathbf{P}, \mathbf{Q}) - d(\mathbf{P}, \mathbf{P})$) comme la « distance » entre $\mathbf{Q}$ et $\mathbf{P}$, de sorte que $\mathbf{P}$ peut être définie comme la valeur de $\mathbf{Q}$ pour laquelle $d(\mathbf{P}, \mathbf{Q})$ atteint son minimum.

Supposons maintenant que $X \in \mathbf{P}$ et que $\mathbf{P}$ est inconnue et appartient à une famille $\mathscr{P}$. Désignons par $(\mathbf{Q})_{\mathscr{P}}$ la distribution de $\mathscr{P}$ la plus proche de $\mathbf{Q}$ au sens de la distance d et supposons qu'elle existe :

$$d((\mathbf{Q})_{\mathscr{P}}, \mathbf{Q}) = \min_{\Pi \in \mathscr{P}} d(\Pi, \mathbf{Q}),$$

si bien que $(\mathbf{Q})_{\mathscr{P}} = \mathbf{Q}$ pour $\mathbf{Q} \in \mathscr{P}$.

DÉFINITION 1. On appelle *estimateur de la distribution* $\mathbf{P}$ *par le minimum de la distance* d la distribution $\mathbf{P}^* = (\mathbf{P}_n^*)_{\mathscr{P}} \in \mathscr{P}$, où $\mathbf{P}_n^*$ est comme toujours une distribution empirique.

Donc, la distance $d(\Pi, \mathbf{P}_n^*)$ est minimale pour $\Pi = \mathbf{P}^* = (\mathbf{P}_n^*)_{\mathscr{P}}$. Si $\mathscr{P}$

est confondue avec l'ensemble de toutes les fonctions de répartition, il est alors évident que $\mathbf{P}^* = \mathbf{P}_n^*$.

Supposons maintenant que $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ est une famille paramétrique vérifiant la condition suivante :

$$(A_0) \qquad \mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2} \quad pour \quad \theta_1 \neq \theta_2.$$

Dans ce cas l'application $\theta \to \mathbf{P}_\theta$ est bijective, de sorte que la distribution $\mathbf{P} \in \mathscr{P}$ permet de déterminer de façon unique le paramètre θ pour lequel $\mathbf{P} = \mathbf{P}_\theta$. Ce fait s'exprime encore comme suit : il existe une fonctionnelle G définie sur $\mathscr{P}$ telle que $\theta = G(\mathbf{P}_\theta)$.

Introduisons la fonctionnelle $G_1(\mathbf{Q}) = G((\mathbf{Q})_{\mathscr{P}})$. Il s'agit de toute évidence de la valeur de $\theta \in \Theta$ pour laquelle $\mathbf{P}_\theta$ sera la distribution la plus proche de $\mathbf{Q}$ au sens de la distance d, de sorte que

$$G_1(\mathbf{P}_\theta) = G(\mathbf{P}_\theta) = \theta. \tag{1}$$

DÉFINITION 2. L'estimateur $\theta^* = G_1(\mathbf{P}_n^*)$ s'appelle *estimateur du paramètre θ par le minimum de la distance d.*

En d'autres termes, l'estimation θ^* est la valeur de Θ pour laquelle

$$d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*) = \inf_{\theta\in\Theta} d(\mathbf{P}_\theta, \mathbf{P}_n^*).$$

Il est évident que nous avons de nouveau affaire au principe de substitution. Ceci résulte des définitions et de (1). Il va de soi que la distance d et la famille $\mathscr{P} = \{\mathbf{P}_\theta\}$ doivent posséder des propriétés assurant la mesurabilité de l'application de $\mathscr{X}^n$ dans R^k, réalisée par la fonctionnelle $G_1(\mathbf{P}_n^*)$, de telle sorte que θ^* soit une variable aléatoire.

Signalons maintenant que *dans le cas paramétrique, si la condition* (A_0) *est remplie, la restriction de la méthode de substitution* (cf. (3.1), (3.2)) *et la méthode de la distance minimale fournissent la même classe d'estimateurs.*

En effet, on sait déjà que les estimateurs θ^* par le minimum de la distance sont des estimateurs de substitution, et de plus $\theta^* \in \Theta$. Supposons maintenant que θ^* est un estimateur de substitution : $\theta^* = G(\mathbf{P}_n^*)$, où $G(\mathbf{P}_\theta) = \theta$, $\theta^* \in \Theta$. Définissons la distance $d(\mathbf{P}, \mathbf{Q}) = |G(\mathbf{P}) - G(\mathbf{Q})|$. Il est alors évident que θ^* réalise

$$\inf_{\theta\in\Theta} d(\mathbf{P}_\theta, \mathbf{P}_n^*) = \inf_{\theta\in\Theta} |G(\mathbf{P}_\theta) - G(\mathbf{P}_n^*)| = \inf_{\theta\in\Theta} |\theta - G(\mathbf{P}_n^*)| = 0.$$

Signalons aussi que le champ d'application de la méthode des moments est bien plus étroit que celui de la méthode de substitution, puisqu'il est évident que les fonctionnelles G telles que $G(\mathbf{P}_\theta) = \theta$ n'admettent pas toutes une représentation de la forme

$$G(\mathbf{P}_\theta) = m^{-1}\left(\int g(x)\mathbf{P}_\theta(dx)\right).$$

Revenons maintenant aux estimateurs par le minimum de la distance. Il est clair qu'on peut indiquer de nombreuses distances « raisonnables » d utilisables pour la construction d'estimateurs. Pour d nous aurions pu prendre la distance

$$d(\mathbf{P}, \mathbf{Q}) = \sup_x |F_{\mathbf{P}}(x) - F_{\mathbf{Q}}(x)|$$

ou

$$d(\mathbf{P}, \mathbf{Q}) = \int (F_{\mathbf{P}}(x) - F_{\mathbf{Q}}(x))^2 dF_{\mathbf{Q}}(x),$$

où $F_{\mathbf{P}}(x)$ est la fonction de répartition de la distribution $\mathbf{P}$. Les estimateurs θ^* par le minimum de la distance seront ici les valeurs de θ qui réalisent respectivement

$$\inf_\theta \sup_x |F_{\mathbf{P}_\theta}(x) - F_n^*(x)|,$$

$$\inf_\theta \int (F_{\mathbf{P}_\theta}(x) - F_n^*(x))^2 dF_n^*(x) = \inf_\theta \frac{1}{n} \sum_{k=1}^{n} \left(F_{\mathbf{P}_\theta}(x_{(k)}) - \frac{k-1}{n}\right)^2. \quad (2)$$

Dans certains problèmes (comparer avec [19]) on utilise les estimateurs par le minimum du χ^2. Il s'agit des estimateurs par le minimum de la distance

$$d(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{r} \frac{(\mathbf{P}(\Delta_i) - \mathbf{Q}(\Delta_i))^2}{\mathbf{P}(\Delta_i)},$$

où $\Delta_1, \ldots, \Delta_r$ est une partition de R (ou de R^m si x_j sont m-dimensionnels) en $r < \infty$ intervalles, de sorte que $\bigcup_{i=1}^{r} \Delta_i = R$. L'estimation θ^* par le minimum du χ^2 est donc la valeur de θ qui minimise

$$n \sum_{i=1}^{r} \frac{(\mathbf{P}_\theta(\Delta_i) - \nu_i/n)^2}{\mathbf{P}_\theta(\Delta_i)} = \sum_{i=1}^{r} \frac{(n\mathbf{P}_\theta(\Delta_i) - \nu_i)^2}{n\mathbf{P}_\theta(\Delta_i)}, \quad (3)$$

où $\nu_i = n\mathbf{P}_n^*(\Delta_i)$ est le nombre d'observations x_j contenues dans l'intervalle Δ_i. Cette estimation tient son nom de la statistique χ^2 qui figure au second membre de (3).

Nous verrons plus bas qu'il existe une fonctionnelle G, telle que $G(\mathbf{P}_\theta) = \theta$ et pour laquelle les estimateurs de substitution appelés estimateurs par le maximum de vraisemblance seront meilleurs dans un certain sens. De ce fait les estimateurs envisagés dans ce paragraphe ne font pas recette dans les applications et ne seront donc pas étudiés en détail.

§ 6. Méthode du maximum de vraisemblance

Soit $\mathscr{P}$ une famille paramétrique de distributions $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$. Dans la suite, chaque fois qu'on aura affaire à cette famille, on admettra que sont réalisées les conditions :

$$(A_0) \qquad \mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2} \quad pour \quad \theta_1 \neq \theta_2,$$

(A_μ) : *Il existe une mesure σ-finie μ sur l'espace mesurable $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ telle que toutes les distributions $\mathbf{P}_\theta \in \mathscr{P}$ admettent la densité $f_\theta(x) = \dfrac{d\mathbf{P}_\theta}{d\mu}(x)$, de sorte que*

$$\mathbf{P}_\theta(B) = \int_B f_\theta(x)\mu(dx).$$

On dit alors que la mesure μ *domine* les distributions $\mathbf{P}_\theta$.

Toutes les familles de distributions étudiées au § 2 vérifient visiblement les conditions (A_0) et (A_μ). Pour μ on prendra la mesure de Lebesgue si les distributions étudiées sont absolument continues et une mesure cardinale, si elles sont discrètes. Par mesure cardinale nous entendons une mesure μ telle que $\mu(B) = k$, où k est le nombre de points de B à coordonnées entières.

Aux premières se rapportent les distributions $\mathbf{\Phi}_{\alpha,\sigma^2}$, $\mathbf{L}_{\alpha,\sigma^2}$, gamma, bêta, uniforme, de Cauchy, de Student et de Ficher. Aux deuxièmes, les distributions de Bernoulli, de Poisson, dégénérée en 0 et polynomiale. Les densités $f_\theta(x)$ de ces distributions sont citées dans le § 2. Dans le cas discret (lorsque μ est une mesure cardinale), la densité $f_\theta(x)$ est confondue avec la probabilité $\mathbf{P}_\theta(\{x\})$ de l'événement $\{x_1 = x\}$; ici $\{x\}$ désigne l'ensemble composé du seul point x. Signalons aussi que les distributions $\mathbf{\Phi}_{\alpha,\sigma^2}$ et de Poisson par exemple sont mutuellement singulières. Au lieu de la mesure de Lebesgue et de la mesure cardinale, on pourrait envisager d'autres mesures, par exemple les distributions $\mathbf{\Phi}_{0,1}$ et $\mathbf{\Pi}_1$ respectivement. Mais les densités $f_\theta(x)$ seraient alors différentes. Nous proposons au lecteur de les déterminer. Les exemples ci-dessus se rapportent au cas où $\mathscr{X} = R$ ou $\mathscr{X} = R^m$, $m > 1$. La mesure μ peut être de nature plus complexe dans un espace des phases arbitraire $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$.

La condition (A_μ) est commode tout d'abord par ce qu'elle nous permet d'envisager, sous un même point de vue, l'étude des deux types de distributions les plus importantes dans les applications : les distributions absolument continues et les distributions discrètes. La condition (A_μ) ne fait pas de distinction entre ces distributions. De plus, la dimension de l'espace des phases $\mathscr{X}$ devient inessentielle.

On conviendra d'écrire

$$f(x) = g(x) \quad [\mu]\text{-presque partout}$$

s'il existe un ensemble μ-negligeable A (i.e. $\mu(A) = 0$) tel que $f(x) = g(x)$ pour tous les $x \notin A$. Il est évident que $f(x) = g(x)$ $[\mu]$-presque partout si et seulement si

$$\int (f(x) - g(x))^2 \mu(dx) = 0.$$

LEMME 1. *Soient f et g deux densités de probabilité par rapport à la mesure μ. Alors*

$$\int f(x) \ln f(x) \mu(dx) \geqslant \int f(x) \ln g(x) \mu(dx) \tag{1}$$

si ces deux intégrales sont finies. Le signe d'égalité n'est possible que si $f =$ $= g$ $[\mu]$-presque partout.

On conviendra que les intégrales de (1), étendues à un ensemble A sur lequel $f(x) = 0$, sont nulles quelle que soit $g(x)$.

DÉMONSTRATION. Nous devons démontrer que

$$\int f(x) \ln \frac{g(x)}{f(x)} \mu(dx) \leqslant 0.$$

Puisque $\ln(1 + x) \leqslant x$ pour tous les $x \geqslant -1$ et que l'égalité n'est possible que pour $x = 0$, il vient

$$\ln \frac{g(x)}{f(x)} = \ln\left(1 + \left(\frac{g(x)}{f(x)} - 1\right)\right) \leqslant \frac{g(x)}{f(x)} - 1,$$

où le signe d'égalité n'est possible que pour $f(x) = g(x)$. Donc,

$$\int f(x) \ln \frac{g(x)}{f(x)} \mu(dx) \leqslant \int f(x)\left(\frac{g(x)}{f(x)} - 1\right)\mu(dx) =$$
$$= \int g(x)\mu(dx) - \int f(x)\mu(dx) = 0. \tag{2}$$

Il est évident que l'inégalité (2) sera stricte si $f = g$ $[\mu]$-presque partout est mise en défaut. ◄

Considérons maintenant une famille $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta \in \Theta}$ satisfaisant les conditions (A_0) et (A_μ), et la distance $d(\mathbf{P}_\theta, \mathbf{Q})$ entre une distribution arbitraire $\mathbf{Q}$ et une distribution $\mathbf{P}_\theta \in \mathscr{P}$

$$d(\mathbf{P}_\theta, \mathbf{Q}) = -\int \ln f_\theta(x) \mathbf{Q}(dx). \tag{3}$$

Définissons la fonctionnelle $G(\mathbf{Q})$ comme la valeur de θ qui réalise

$$\min_\theta d(\mathbf{P}_\theta, \mathbf{Q}) = d(\mathbf{P}_{G(\mathbf{Q})}, \mathbf{Q}).$$

Le lemme 1 et la condition (A_0) entraînent

$$-\int f_{\theta_0} \ln f_\theta \mu(dx) > -\int f_{\theta_0} \ln f_{\theta_0}\mu(dx),$$
$$d(\mathbf{P}_\theta, \mathbf{P}_{\theta_0}) > d(\mathbf{P}_{\theta_0}, \mathbf{P}_{\theta_0})$$

pour $\theta \neq \theta_0$. Ce qui exprime que

$$G(\mathbf{P}_{\theta_0}) = \theta_0. \tag{4}$$

DÉFINITION 1. On appelle *estimation par le maximum de vraisemblance* l'estimation $\hat{\theta}^* = G(\mathbf{P}_n^*)$, c'est-à-dire la valeur de θ qui réalise

$$\max_\theta \int \ln f_\theta(x)\mathbf{P}_n^*(dx) = \max_\theta \frac{1}{n}\sum_{i=1}^{n} \ln f_\theta(x_i). \tag{5}$$

Le symbole ˆ désignera dans la suite les estimateurs par la méthode du maximum de vraisemblance.

De la définition et de (4) il résulte que *l'estimateur par le maximum de vraisemblance est un estimateur de substitution.* On peut le traiter aussi comme un estimateur par le minimum de la distance (3). Cette distance est étroitement liée à la distance de Kullback-Leibler entre les distributions. Cette distance qui joue un rôle particulier en statistique mathématique sera examinée ultérieurement.

Dans la définition 1, la famille $\{\mathbf{P}_\theta\}$ est supposée telle que $\hat{\theta}^*$ soit une variable aléatoire *).

L'estimateur par le maximum de vraisemblance n'est pas unique puisqu'une fonction peut atteindre son maximum en plusieurs points. Nous citerons un exemple plus bas.

La dénomination de cet estimateur est liée à l'importante interprétation suivante de l'expression

$$\sum_{i=1}^{n} \ln f_\theta(x_i) = \ln \prod_{i=1}^{n} f_\theta(x_i)$$

qui figure dans (5). Pour simplifier, considérons tout d'abord le cas discret où μ est une mesure cardinale. Alors $\prod_{i=1}^{n} f_\theta(x_i)$ est la probabilité d'apparition de l'issue $X = (x_1, \ldots, x_n)$. Nous prenons donc pour estimation $\hat{\theta}^*$ la valeur du paramètre θ qui *maximise cette probabilité* (en effet, les fonctions $\varphi(\theta) > 0$ et $\ln \varphi(\theta)$ atteignent leurs extremums aux mêmes points).

---

*) Autrement dit, $\hat{\theta}^*$ est une application mesurable de $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ dans $(R^k, \mathfrak{B}^k)$.

Cette interprétation est valable pour le cas général aussi. Les x_i étant indépendants, on a pour les ensembles $B = B_1 \times \ldots \times B_n$, $B_i \in \mathfrak{B}_{\mathscr{X}}$,

$$\mathbf{P}_\theta(X \in B) = \int_{B_1} f_\theta(x_1)\mu(dx_1) \ldots \int_{B_n} f_\theta(x_n)\mu(dx_n). \tag{6}$$

On rappelle que x_i sont des variables et que le vecteur $(x_1, \ldots, x_n)$ est désigné par $\mathbf{x}$. Soit μ^n le n-uple produit direct des mesures μ, de sorte que $\mu^n(d\mathbf{x}) = \prod_{i=1}^{n} \mu(dx_i)$. Alors (6) exprime que

$$\mathbf{P}_\theta(X \in B) = \int_B \Big(\prod_{i=1}^{n} f_\theta(x_i)\Big)\mu^n(d\mathbf{x})$$

et donc que la fonction $f_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i)$ est la densité de probabilité du vecteur aléatoire X dans $\mathscr{X}^n$ par rapport à la mesure μ^n,

$$\int_{\mathscr{X}^n} f_\theta(\mathbf{x})\mu^n(d\mathbf{x}) = 1.$$

Donc, $\prod_{i=1}^{n} f_\theta(x_i)\mu^n(d\mathbf{x})$ peut être interprété (par analogie au cas discret) comme la probabilité que l'échantillon tombe dans le parallélépipède formé par l'intersection des « bandes » $]x_i, x_i + dx_i[$, et l'estimateur par le maximum de vraisemblance maximise cette probabilité par rapport à θ.

La fonction

$$f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i)$$

traitée comme une fonction de θ s'appelle *fonction de vraisemblance*, et la fonction

$$L(X, \theta) = \ln f_\theta(X) = \sum_{i=1}^{n} l(x_i, \theta),$$

où $l(x, \theta) = \ln f_\theta(x)$, *logarithme de la fonction de vraisemblance*.

Nous réserverons ces noms aux fonctions f et L dans le cas où l'argument est le vecteur $\mathbf{x}$. Donc, la fonction de vraisemblance $f_\theta(\mathbf{x})$ est une fonction sur $\mathscr{X}^n \times \Theta$, qui pour chaque $\theta \in \Theta$ est une densité de probabilité par rapport à la mesure μ^n, de sorte que la densité $f_\theta(x_1)$ dans $\mathscr{X}$ est aussi une fonction de vraisemblance pour $n = 1$.

D'autre part, pour $\mathscr{X} = R$ par exemple, la fonction $f_\theta(X)$ peut être traitée comme une fonction de vraisemblance d'un échantillon de taille 1 dans le cas multidimensionnel où $\mathscr{X} = R^m = R^n$.

Il est important de souligner que l'estimateur par le maximum de vraisemblance est totalement indépendant du choix de μ, puisque la substitu-

tion à μ d'une mesure équivalente μ_1 se traduit par la multiplication de la fonction de vraisemblance $f_\theta(x)$ par un facteur $\frac{d\mu^n}{d\mu_1^n}(x)$ indépendant de θ.

Pour étudier les *propriétés asymptotiques des estimateurs par le maximum de vraisemblance*, on pourrait suivre la même marche que pour les estimateurs par la méthode des moments. On rappelle qu'on avait utilisé le fait que les estimateurs par la méthode des moments sont des statistiques du premier type. Ceci nous a permis d'établir immédiatement leur convergence forte et leur normalité asymptotique. Sous certaines conditions sur $f_\theta(x)$ *les estimateurs par le maximum de vraisemblance seront des statistiques du deuxième type*, ce qui nous permettra (cf. théorèmes des §§ 1.5 et 1.8) d'établir leur convergence et leur normalité asymptotique. Mais nous étudierons les propriétés des estimateurs par le maximum de vraisemblance directement (cf. §§ 23 à 27) pour des raisons d'économie et d'exhaustivité.

Trouvons les fonctions de vraisemblance et les estimateurs par le maximum de vraisemblance pour certaines distributions du § 2. Si les fonctions de vraisemblance sont régulières, il est plus simple de déterminer leur maximum en égalant les dérivées premières à zéro.

EXEMPLE 1. La distribution normale Φ_{α,σ^2} dans $\mathscr{X} = R$ admet la densité

$$\varphi_{\alpha,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}, \quad -\infty < \alpha < \infty, \quad \sigma > 0.$$

En supposant que $\theta^* = (\alpha, \sigma^2)$, on obtient

$$f_\theta(x) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \alpha)^2\right\},$$

$$L(X, \theta) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \alpha)^2.$$

Comme déjà signalé, les fonctions f et L atteignent leurs maximums pour les mêmes θ, puisque ln est monotone. On a

$$\frac{\partial L}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \alpha),$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \alpha)^2.$$

La résolution du système d'équations

$$\frac{\partial L}{\partial \alpha} = 0, \quad \frac{\partial L}{\partial \sigma} = 0$$

nous donne

$$\hat{\alpha}^* = \bar{x}, \quad (\hat{\sigma^2})^* = S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Il est immédiat de vérifier que L atteint bien son maximum en ce point.

EXEMPLE 2. Considérons une distribution gamma de densité

$$\gamma_\alpha(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, \quad x \geqslant 0, \quad \alpha > 0,$$

dans le cas où le paramètre λ est connu. On a

$$L(X, \alpha) = \lambda n \ln\alpha - n \ln\Gamma(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i - \alpha \sum_{i=1}^{n} x_i,$$

$$\frac{\partial L}{\partial \alpha} = \frac{\lambda_n}{\alpha} - \bar{x}n, \quad \hat{\alpha}^* = \lambda/\bar{x}.$$

EXEMPLE 3. Distribution binomiale $\mathbf{B}_p$. Pour $X \in \mathbf{B}_p$, on a $\mathbf{P}(x_i = 1) = p$, $\mathbf{P}(x_i = 0) = 1 - p$,

$$f_p(X) = p^\nu(1 - p)^{n-\nu},$$

où ν est le nombre d'apparitions de 1 parmi les éléments $x_1, \ldots, x_n$. Donc,

$$L(X, p) = \nu \ln p + (n - \nu) \ln (1 - p),$$

$$\frac{\partial L}{\partial p} = \frac{\nu}{p} - \frac{n - \nu}{1 - p}, \quad \hat{p}^* = \frac{\nu}{n}.$$

Nous proposons au lecteur de trouver à titre d'exercice les estimateurs par le maximum de vraisemblance de toutes les familles paramétriques du § 2 et de les comparer à ceux de la méthode des moments.

Citons maintenant deux exemples dans lesquels la fonction f_θ n'est pas régulière par rapport à θ et les méthodes de recherche d'un estimateur par le maximum de vraisemblance impliquant une dérivation ne passent pas.

EXEMPLE 4. Soit $X \in \mathbf{U}_{\theta, 1+\theta}$ (la distribution uniforme sur $[\theta, 1 + \theta]$). On a

$$f_\theta(x) = \begin{cases} 1 & \text{si } x \in [\theta, 1 + \theta], \\ 0 & \text{sinon,} \end{cases}$$

$$f_\theta(X) = \begin{cases} 1 & \text{si } \theta \leqslant x_{(1)} \leqslant x_{(n)} \leqslant 1 + \theta, \\ 0 & \text{sinon,} \end{cases}$$

où $x_{(1)} \leqslant \ldots \leqslant x_{(n)}$ est un échantillon ordonné. L'estimateur par le maximum de vraisemblance n'est pas unique dans cet exemple. En effet, $f_\theta(X) = 1$ (c'est-à-dire à la valeur maximale) pour tout θ vérifiant la double inégalité $x_{(n)} - 1 \leqslant \theta \leqslant x_1$. De tels θ existent toujours puisque $x_{(n)} - x_{(1)} \leqslant 1$. On peut prendre en particulier $\hat{\theta}^* = x_{(1)}$ ou $\hat{\theta}^* = x_{(n)} - 1$.

EXEMPLE 5. Soit $X \in U_{0,\,\theta}$. On a

$$f_\theta(x) = \begin{cases} \theta^{-1} & \text{si } x \in [0, \theta], \\ 0 & \text{sinon,} \end{cases}$$

$$f_\theta(X) = \begin{cases} \theta^{-n} & \text{si } x_i \in [0, \theta] \text{ pour } i = 1, \ldots, n, \\ 0 & \text{sinon.} \end{cases}$$

Pour exprimer $f_\theta(X)$ comme une fonction de θ, écrivons la condition $x_i \in [0, \theta]$, $i = 1, \ldots, n$, sous la forme équivalente $\theta \geqslant \max x_i = x_{(n)}$. Donc, $f_\theta(X) = 0$ pour $\theta \in [0, x_{(n)}[$, et $f_\theta(X) = \theta^{-n}$ pour $\theta \in]x_{(n)}, \infty[$. Le graphique de cette fonction est représenté sur la figure 1. La fonction f_θ est discontinue comme dans l'exemple précédent. Elle atteint son maximum au point $\hat{\theta}^* = x_{(n)}$.

Le lecteur trouvera de façon analogue un estimateur par le maximum de vraisemblance du paramètre inconnu (α, β) lorsque $X \in U_{\alpha,\,\beta}$.

Si $f_\theta(x)$ est infinie en des points x_θ dépendant de θ, la méthode du maximum de vraisemblance n'a plus de sens (nous avons convenu ici que $f_\theta(x_\theta) = \infty$ si $f_\theta(x) \to \infty$ lorsque $x \downarrow x_\theta$ ou $x \uparrow x_\theta$). Ceci est plus facile à comprendre sur l'exemple du paramètre de translation lorsque $f_\theta(x) = f(x - \theta)$, $f(x) > 0$,

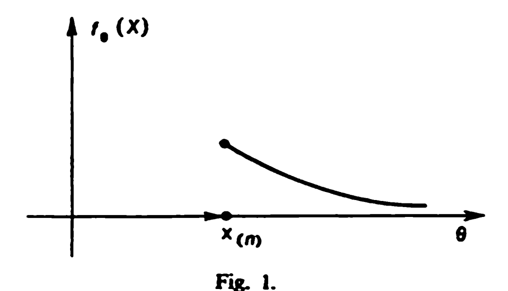


Fig. 1.

$f(0) = \infty$. Dans ce cas $f_\theta(X) = \infty$ pour $\theta = x_1, \ldots, \theta = x_n$ et par suite $\hat{\theta}^*$ prend au moins n valeurs confondues avec les éléments de l'échantillon. L'explication de cet effet est que les « jaillissements » de la fonction $f_\theta(X)$ ne permettent pas de juger de la position du « vrai » maximum de $f_\theta(X)$ qui est conditionné par l'échantillon (comparer avec les §§ 24 et 25). Pour pouvoir le faire il faudrait « lisser » par un quelconque procédé la fonction $f_\theta(X)$.

Les estimateurs par le maximum de vraisemblance jouissent de l'importante propriété suivante d'invariance par le changement du paramètre.

THÉORÈME 1. *Soit $\beta(\theta)$ une application bijective de l'ensemble Θ sur l'ensemble B. Si $\hat{\theta}^*$ est un estimateur du maximum de vraisemblance du paramètre θ, alors $\hat{\beta}^* = \beta(\hat{\theta}^*)$ sera un estimateur du maximum de vraisemblance du paramètre $\beta = \beta(\theta)$ pour une famille paramétrique $\{\mathbf{Q}_\beta = \mathbf{P}_{\theta(\beta)}\}_{\beta \in B}$, où $\theta(\beta)$ est la fonction réciproque de $\beta(\theta)$.*

On glissera sur la démonstration de ce théorème, car elle coule de source.

Signalons que nous avons déjà implicitement utilisé le théorème 1 dans l'exemple 1 où pour déterminer l'estimateur du maximum de vraisemblance de σ^2 nous avons cherché le maximum de L par rapport à σ et ensuite avons pris $(\hat{\sigma^2})^* = (\hat{\sigma}^*)^2$.

Un autre exemple d'application de ce théorème est la recherche de l'estimateur par le maximum de vraisemblance pour $X \in \mathbf{L}_{\alpha,\sigma^2}$, c'est-à-dire dans le cas où $\ln x_i \in \Phi_{\alpha,\sigma^2}$. La moyenne a et la variance d^2 de tels x_i sont respectivement égales à (cf. § 2)

$$a = \exp\{\alpha + \sigma^2/2\}, \quad d^2 = a^2(e^{\sigma^2} - 1).$$

Si l'on désigne les estimateurs du maximum de vraisemblance de a et de d^2 respectivement par $\hat{a}^*$ et $(\hat{d^2})^*$, on obtient pour la fonction $(a, d^2) = \beta(\alpha, \sigma^2)$ (cf. exemple 1)

$$\hat{a}^* = \exp\left\{\bar{y} + \frac{S_Y^2}{2}\right\}, \quad (\hat{d^2})^* = (\hat{a}^*)^2(e^{S_Y^2} - 1),$$

où

$$Y = (y_1, \ldots, y_n), \quad y_i = \ln x_i, \quad \bar{y} = \frac{1}{2}\sum_{i=1}^{n} y_i, \; S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Le § 26 traite du calcul approché de l'estimateur par le maximum de vraisemblance dans des situations plus générales.

Avant de refermer ce paragraphe faisons la remarque suivante. Nous avons déjà dit que l'estimateur du maximum de vraisemblance était un estimateur de substitution. Mais l'estimateur du maximum de vraisemblance peut être traité, sous certaines conditions, comme un estimateur de la méthodes des moments généralisée. Supposons en effet que la fonction $f_\theta(x)$ est dérivable par rapport à θ et que cette dérivation est licite sous le signe d'intégration dans l'égalité

$$\int f_\theta(x)\mu(dx) = 1.$$

7—4195

Alors

$$0 = \int f'_\theta(x)\mu(dx) = \int\limits_{\{f_\theta(x)\neq 0\}} \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x)\mu(dx) = \int\limits_{\{f_\theta(x)\neq 0\}} l'(x, \theta)f_\theta(x)\mu(dx) = \mathbf{E}_\theta l'(x_1, \theta).$$

Si donc dans (4.6) on pose $g(x, \theta) = l'(x, \theta)$, on obtient pour l'estimateur par la méthode des moments généralisée l'équation

$$\int l'(x, \theta)\mathbf{P}_n^*(dx) = \int l'(x, \theta)\mathbf{P}_\theta(dx) = 0$$

ou ce qui est équivalent

$$L'(X, \theta) = 0.$$

Ceci est l'équation pour l'estimateur par la méthode du maximum de vraisemblance.

§ 7. Sur la comparaison des estimateurs

Nous avons vu qu'il existait plusieurs méthodes assez naturelles de construction des estimateurs. Une question se pose : comment comparer ces estimateurs et quels sont les meilleurs ? A cet effet on dispose de deux approches : l'approche de la moyenne quadratique et l'approche asymptotique.

La première repose sur la comparaison des dispersions quadratiques moyennes. La seconde ne s'applique qu'aux échantillons de grande taille, car elle s'appuie sur la comparaison des « dispersions » des distributions de $(\theta^* - \theta)\sqrt{n}$ pour de grands n. Cette comparaison est généralement basée sur la forme des distributions limites de $(\theta^* - \theta)\sqrt{n}$ (si elles existent) lorsque $n \to \infty$. Les théorèmes limites correspondants nous donnent les conditions sous lesquelles la distribution de $(\theta^* - \theta)\sqrt{n}$ pour les grands n peut être approchée par les distributions limites mentionnées.

Dans ce paragraphe, on admet que les estimateurs sont comparés pour une distribution inconnue mais fixée $\mathbf{P}$.

1. Approche de la moyenne quadratique. Cas scalaire. Cette approche est utilisée pour étudier les estimateurs au vu d'un échantillon X de taille quelconque fixée (non nécessairement élevée). Elle consiste à comparer les erreurs quadratiques moyennes $\mathbf{E}(\theta^* - \theta)^2$.

RÈGLE 1. *On dira qu'un estimateur θ_1^* est meilleur en moyenne quadratique qu'un estimateur θ_2^* si*

$$\mathbf{E}(\theta_1^* - \theta)^2 < \mathbf{E}(\theta_2^* - \theta)^2. \tag{1}$$

L'idée que l'erreur quadratique moyenne est la caractéristique numérique la mieux appropriée de précision d'une estimation est largement répandue bien qu'elle soit discutable à de nombreux égards : en effet, on peut comparer par exemple les quantités $\mathbf{E}|\theta^* - \theta|$ qui, elles aussi, décrivent les écarts moyens entre θ^* et θ.

L'avantage indéniable de $\mathbf{E}(\theta^* - \theta)^2$ est que $(\theta^* - \theta)^2$ est une fonction

analytique de la différence $\theta^* - \theta$. Ceci rend l'étude plus commode et permet, comme nous le verrons plus bas, d'approcher les valeurs de $\mathbf{E}f(\theta^* - -\theta)$ pour des fonctions régulières f.

Pour décrire les propriétés des estimateurs, on se sert aussi du biais.

DÉFINITION 1. On appelle *biais* ou *erreur systématique* d'un estimateur θ^* la quantité

$$b = \mathbf{E}\theta^* - \theta.$$

Un estimateur θ^* pour lequel $b = 0$ est dit *sans biais* ou *non biaisé*.

Entre l'erreur quadratique moyenne, le biais et la variance d'un estimateur, on a la relation

$$\mathbf{E}(\theta^* - \theta)^2 = \mathbf{V}\theta^* + b^2,$$

de sorte que l'erreur quadratique moyenne et la variance des estimateurs sans biais coïncident.

La propriété d'être sans biais est visiblement souhaitable, car elle exprime que dans une suite d'estimations donnée, la moyenne des estimations sera confondue avec la vraie valeur du paramètre. Si cette propriété fait défaut, on dit que l'estimation est *à biais* ou *biaisée*.

EXEMPLE 1. Considérons les trois estimateurs suivants de la valeur moyenne $\theta = \mathbf{E}x_1$ d'une distribution $\mathbf{P}$:

$$\theta_1^* = \bar{x}, \quad \theta_2^* = \zeta^*, \quad \theta_3^* = \frac{x_{(1)} + x_{(n)}}{2}, \tag{2}$$

où ζ^* est la médiane empirique, $x_{(k)}$, $k = 1, \ldots, n$, les éléments de l'échantillon ordonné, de sorte que $\zeta^* = x_{((n+1)/2)}$ si n est impair et $\zeta^* = \frac{1}{2}(x_{(n/2)} + + x_{(n/2+1)})$ si n est pair (les trois estimateurs sont confondus pour $n = 1, 2$). Tous ces estimateurs sont sans biais si la distribution $\mathbf{P}$ est symétrique par rapport à θ : $\mathbf{P}(]-\infty, \theta - t[) = \mathbf{P}(]\theta + t, \infty[)$, $\forall t \geqslant 0$. Ceci résulte du fait que la distribution de ces trois estimateurs sera aussi symétrique par rapport à θ. Il est évident que $\mathbf{E}\bar{x} = \theta$ est sans biais sans l'hypothèse de symétrie.

Calculons les erreurs quadratiques moyennes des estimateurs (2). Pour simplifier on se bornera au cas $\mathbf{P} = \mathbf{U}_{0,1}$, $n = 3$. Les estimateurs (2) deviennent

$$\theta_1^* = \bar{x}, \ \theta_2^* = x_{(2)}, \quad \theta_3^* = \frac{x_{(1)} + x_{(3)}}{2}.$$

On a

$$\mathbf{V}x_1 = \int_0^1 (x - 1/2)^2 dx = 1/12,$$

$$\mathrm{E}(\theta_1^* - \theta)^2 = \mathrm{V}\bar{\mathrm{x}} = \mathrm{Vx}_1/3 = 1/36.$$

Par définition de la médiane (n est impair), $\{\zeta^* < x\} = \{F_n^*(x) > 1/2\}$, donc

$$\mathrm{P}(\zeta^* < x) = \mathrm{P}(F_n^*(x) > 1/2) = \sum_{k=(n+1)/2}^{n} \mathrm{P}(nF_n^*(x) = k). \tag{3}$$

Pour $n = 3$

$$\mathrm{P}(F_3^*(x) = 1) = \mathrm{P}\left(\bigcap_{i=1}^{3} \{\mathrm{x}_i < x\}\right) = F^3(x),$$

$$\mathrm{P}(3F_3^*(x) = 2) = 3F^2(x)(1 - F(x)).$$

La probabilité $\mathrm{P}(\zeta^* \in]u, u + du[)$ est composée des probabilités d'événements de la forme $\{\mathrm{x}_1 \in]u, u + du[\}\{\mathrm{x}_2 < u\}\{\mathrm{x}_3 > u\}$. Ces combinaisons étant au nombre de 6, il vient $\mathrm{P}(\zeta^* \in]u, u + du[) = 6f(u)F(u)(1 - F(u))du$ et par suite ζ^* admet une densité égale à

$$6f(u)F(u)(1 - F(u)),$$

où $F(u) = \int_{-\infty}^{u} f(t)dt = \mathrm{P}(\mathrm{x}_1 < u)$ (ceci résulte aussi de (3)). Si $\mathbf{P} = \mathrm{U}_{0,1}$, cette densité sera égale à $6x(1 - x)$, $x \in [0, 1]$, de sorte que

$$\mathrm{E}(\zeta^*)^2 = \int_0^1 6x^3(1 - x)dx = 6\left(\frac{1}{4} - \frac{1}{5}\right) = \frac{3}{10},$$

$$\mathrm{V}\zeta^* = \mathrm{E}(\zeta^*)^2 - (\mathrm{E}\zeta^*)^2 = \frac{3}{10} - \frac{1}{4} = \frac{1}{20}.$$

Reste à trouver la variance de l'estimateur

$$\theta_3^* = \frac{\mathrm{x}_{(1)} + \mathrm{x}_{(3)}}{2}.$$

En raisonnant comme plus haut, on trouve immédiatement que pour $u < v$ la probabilité $\mathrm{P}(\mathrm{x}_1 \in]u, u + du[, \mathrm{x}_{(3)} \in]v, v + dv[) = 6f(u)f(v)(F(v) - F(u))dudv$. Donc, si $\mathbf{P} = \mathrm{U}_{0,1}$

$$\mathrm{E}(\theta_3^*)^2 = \int_0^1\int_0^v \left(\frac{u + v}{2}\right)^2 6(v - u)du\,dv.$$

Cette intégrale est égale à 11/40 (le détail des calculs est laissé au soin du lecteur), de sorte que

$$\mathrm{V}\theta_3^* = \mathrm{E}(\theta_3^*)^2 - (\mathrm{E}\theta_3^*)^2 = \frac{11}{40} - \frac{1}{4} = \frac{1}{40}.$$

L'estimateur θ_3^* est donc le meilleur. La situation peut être différente pour les autres valeurs de n et d'autres distributions **P**. Nous verrons par exemple que pour $\mathbf{P} = \Phi_{\alpha,\sigma^2}$ le meilleur estimateur de α sera $\theta_1^* = \bar{x}$.

EXEMPLE 2. *Estimateurs sans biais de la variance.* Considérons les estimateurs (de substitution) de la variance

$$S^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum x_i^2 - (\bar{x})^2,$$

$$S_1^2 = \frac{1}{n}\sum(x_i - \mathbf{E}x_i)^2 = \frac{1}{n}\sum x_i^2 + (\mathbf{E}x_1)^2 - 2\bar{x}\mathbf{E}x_1$$

dans le cas où $\mathbf{E}x_1$ est connue. L'estimateur S_1^2 est visiblement sans biais. Dans le même temps

$$S^2 = \frac{1}{n}\sum(x_i - \bar{x})^2 = \frac{1}{n}\sum(x_i - \bar{x} \pm \mathbf{E}x_1)^2 =$$

$$= \frac{1}{n}\sum(x_i - \mathbf{E}x_1)^2 - (\bar{x} - \mathbf{E}x_1)^2 = S_1^2 - (\bar{x} - \mathbf{E}x_1)^2 < S_1^2.$$

Donc, l'estimateur S^2 est à biais, et

$$\mathbf{E}S^2 = \mathbf{V}x_1 - \mathbf{V}\bar{x} = \left(1 - \frac{1}{n}\right)\mathbf{V}x_1.$$

Cette relation montre que dans le cas aussi où $\mathbf{E}x_1$ est inconnue, on peut considérer un estimateur sans biais de la variance égal à

$$S_0^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2, \quad \mathbf{E}S_0^2 = \mathbf{V}x_1.$$

Passons maintenant à l'approche asymptotique de la comparaison des estimateurs. Dans ce cas nous sommes placés devant un choix unique de l'estimateur.

2. **Approche asymptotique. Cas scalaire.** Soient donnés deux estimateurs θ_1^* et θ_2^* tels que

$$\frac{(\theta_1^* - \theta)\sqrt{n}}{\sigma_1} \in \mathbf{Q}, \quad \frac{(\theta_2^* - \theta)\sqrt{n}}{\sigma_2} \in \mathbf{Q}, \tag{4}$$

où **Q** est une loi limite, la même pour θ_1^* et θ_2^*, et $\sigma_2 > \sigma_1$. Pour les grands n les distributions de $(\theta_i^* - \theta)\sqrt{n}/\sigma_i$, $i = 1, 2$, seront alors proches de **Q** et la dispersion de θ_2^* autour de θ sera indiscutablement supérieure à celle de θ_1^*. Il nous faut par conséquent préférer θ_1^*.

L'approche asymptotique consiste donc à comparer les distributions limites des estimateurs.

Nous avons déjà vu, et nous nous en assurerons dans la suite, que de nombreux estimateurs, y compris les estimateurs optimaux, sont asymptotiquement normaux, c'est-à-dire sont justiciables de (4) pour $\mathbf{Q} = \Phi_{0,1}$. Ceci nous permet de formuler la règle naturelle suivante de comparaison des estimateurs asymptotiquement normaux.

Soient donnés deux estimateurs asymptotiquement normaux θ_1^* et θ_2^* de paramètres σ_1^2 et σ_2^2 respectivement.

RÈGLE 2. *On dira que l'estimateur θ_1^* est meilleur que θ_2^* si $\sigma_1^2 < \sigma_2^2$.*

Dans la suite, on se servira aussi des termes « aussi bon », « pire » et « pas meilleur » qui correspondront aux signes d'inégalité $\leqslant$, $>$ et $\geqslant$ entre σ_1^2 et σ_2^2 (ou entre $\mathbf{E}(\theta_1^* - \theta)^2$ et $\mathbf{E}(\theta_2^* - \theta)^2$ dans (1)). Si $\sigma_1^2 = \sigma_2^2$, les estimateurs seront dits *asymptotiquement équivalents*. Cette convention est naturelle et dans la suite nous ne la spécifierons pas à chaque fois, nous contentant seulement de définir la relation de « meilleur » ou les relations analogues.

Signalons que dans la classe des estimateurs asymptotiquement normaux, dire que la dispersion de θ^* est minimale revient à dire que la quantité

$$\lim_{n \to \infty} \mathbf{P}(|\theta^* - \theta| < u/\sqrt{n})$$

sera maximale pour tout u. Cette circonstance rend irréprochable cette règle de comparaison des estimateurs asymptotiquement normaux.

En dépit de son caractère naturel, l'approche asymptotique possède un grave défaut : *elle n'est valable que pour les échantillons de grande taille et seulement dans la classe des estimateurs asymptotiquement normaux.*

Les deux approches mentionnées sont dans un certain sens proches l'une de l'autre : dans les deux cas le problème revient à comparer des variances ou des quantités proches des variances. Certes, la quantité σ_i^2/n de (4) pour $\mathbf{Q} = \Phi_{0,1}$ peut se distinguer fondamentalement de $\mathbf{E}(\theta^* - \theta)^2$. Mais les exemples illustrant ce fait (nous proposons au lecteur de les construire) revêtent généralement un caractère artificiel.

La suite de l'exposé de ce chapitre est essentiellement liée à la construction d'estimateurs optimaux pour chacune des deux approches.

EXEMPLE 3. Soit $X \in \Gamma_{\alpha,1}$. Dans l'exemple 1 du § 4, on a montré que les deux estimateurs

$$\alpha_1^* = (\bar{x})^{-1} \quad \text{et} \quad \alpha_2^* = \left(\frac{1}{2n}\sum x_i^2\right)^{-1/2}$$

étaient des estimateurs par la méthode des moments. De plus, α_1^* est en même temps un estimateur par le maximum de vraisemblance. On a établi par ailleurs que ces estimateurs étaient tous deux asymptotiquement normaux de paramètres α^2 et $\frac{5}{4}\alpha^2$ respectivement, de sorte que l'estimateur α_1^*

est meilleur que l'estimateur α_2^* du point de vue de l'approche asymptotique. On obtient le même résultat pour $n \geqslant 2$ dans le cas de l'approche de la moyenne quadratique.

Citons maintenant un exemple montrant qu'un estimateur peut être meilleur ou pire qu'un autre selon les propriétés de la distribution.

EXEMPLE 4. Soit à estimer $\theta = \mathbf{E}x_1$ sachant que $X \in \mathbf{P}$ et $\mathbf{P}$ est symétrique par rapport à θ (comparer avec l'exemple 1). Dans ce cas la médiane ζ de la distribution $\mathbf{P}$ est confondue avec θ. Considérons les deux estimateurs (de substitution) suivants de θ : la moyenne $\theta_1^* = \bar{x}$ et la médiane empirique $\theta_2^* = \zeta^*$. Supposons pour fixer les idées que n est impair. Du corollaire 2.2.1 pour $k = (n + 1)/2$, il s'ensuit que si la fonction de répartition F est continûment dérivable au point $\theta = \zeta$, alors

$$(\zeta^* - \zeta)\sqrt{n} \Rightarrow \frac{\xi}{2f(\theta)}, \quad \xi \in \Phi_{0,\,1}, \quad f(x) = F'(x).$$

En d'autres termes, ζ^* est dans ce cas un estimateur asymptotiquement normal de paramètre $\sigma_2^2 = 1/(4f^2(\zeta))$.

D'autre part, l'estimateur asymptotiquement normal $\bar{x}$ a pour coefficient $\sigma_1^2 = \mathbf{V}x_1$. Donc, si

$$\int (x - \zeta)^2 dF(x) < \frac{1}{4f^2(\zeta)},$$

l'estimateur $\bar{x}$ est meilleur. Si l'inégalité est de sens contraire, c'est ζ^*. Signalons que les nombres $\int (x - \zeta)^2 dF(x)$ et $f(\zeta)$ sont des caractéristiques de la distribution très peu liées entre elles.

Considérons un cas particulier important d'estimateur du paramètre α au vu d'un échantillon $X \in \Phi_{\alpha,\sigma^2}$. Dans ce cas $f(\alpha) = f(\zeta) = \dfrac{1}{\sigma\sqrt{2\pi}}$, de sorte que

$$\sigma_2^2 = \frac{\pi}{2}\sigma^2 > \sigma^2 = \sigma_1^2.$$

Ceci exprime que la statistique $\bar{x}$ est meilleure que ζ^*. Cependant nous avons vu qu'il était aisé de construire un exemple de distribution pour laquelle la statistique ζ^* serait meilleure.

L'exemple de la médiane est instructif à un autre égard. Il montre que le degré de dispersion de $\zeta^* - \zeta$ peut décroître à n'importe quelle vitesse. Pour s'en assurer il suffit de se reporter à la remarque 2.2.1. Dans les conditions de cette remarque, le facteur de normalisation qui est responsable de la convergence de la distribution de $\zeta^* - \zeta$ vers la distribution limite est la quantité $n^{1/(2\gamma)}$, où γ est un nombre positif quelconque (cf. (2.12)). Le facteur $\sqrt{n}$ ne correspond qu'aux distributions régulières.

Citons maintenant une expérience réalisée sur un échantillon de taille $n = 101$ de distribution normale $\Phi_{0,1}$ et voyons *) comment les valeurs $\bar{x}$ et ζ^* tendent vers 0 pour $n = 11, 21, 51, 101$. Les résultats obtenus sont consignés dans le tableau suivant :

n	11	21	51	101
$\bar{x}$ ζ^*	−0,283 −0,291	−0,254 −0,292	−0,148 −0,078	−0,072 −0,044

Dans cet exemple, l'estimateur ζ^* est meilleur pour $n = 51$ et 101, ce qui est la conséquence d'un écart aléatoire. Pour mettre en évidence d'avantage de $\bar{x}$, il aurait fallu réaliser plusieurs expériences de cette nature.

Voyons maintenant quelles formes prennent les deux approches développées dans le cas multidimensionnel, c'est-à-dire lorsque $\theta = (\theta_1, \ldots, \theta_k)$.

3. Approches asymptotique et de la moyenne quadratique dans le cas vectoriel. L'approche asymptotique ne sera utilisée comme toujours que dans la classe des estimateurs asymptotiquement normaux. Dans ce cas le problème se ramène à la comparaison de distributions normales multidimensionnelles (limites pour $(\theta^* - \theta)\sqrt{n}$) qui sont entièrement décrites par la matrice des moments d'ordre deux σ^2 (cf. par exemple théorème 3.2A).

Si l'on se place dans le cadre de l'approche de la moyenne quadratique pour comparer les distributions exactes de θ^*, on aura à comparer deux distributions dans R^k d'après les moments de $(\theta^* - \theta)$ d'ordre deux. Dans les deux cas nous devons donc comparer des matrices des moments d'ordre deux d'après le « degré de dispersion ».

Considérons les procédés de comparaison les plus naturels. Soient $\mathbf{Q}_1$ et $\mathbf{Q}_2$ deux distributions quelconques dans R^k. Soient ξ_1 et ξ_2 des vecteurs aléatoires quelconques de distributions $\mathbf{Q}_1$ et $\mathbf{Q}_2$ respectivement.

DÉFINITION 2. On dira que la dispersion quadratique moyenne de la distribution $\mathbf{Q}_1$ autour d'un point $\alpha \in R^k$ est inférieure à celle de $\mathbf{Q}_2$ si pour tout vecteur $a = (a_1, \ldots, a_k)$

$$\mathbf{E}(\xi_1 - \alpha, a)^2 \leqslant \mathbf{E}(\xi_2 - \alpha, a)^2, \tag{5}$$

où $(x, a) = \sum_{i=1}^{k} x_i a_i$ est le produit scalaire.

On dira que la dispersion quadratique moyenne de $\mathbf{Q}_1$ est strictement inférieure à celle de $\mathbf{Q}_2$ si dans (5) l'inégalité stricte est réalisée au moins pour un a.

---

*) L'échantillon X a été construit à l'aide de nombres aléatoires empruntés aux tableaux de [8].

Si $\alpha = \mathbf{E}\xi_1 = \mathbf{E}\xi_2$, l'inégalité (5) exprime que la variance de Q_1 dans toute direction a (c'est-à-dire la variance de la projection de ξ_1 sur a) est inférieure à la quantité respective de Q_2.

Si $d_l^2 = \| d_{ij}^{(l)} \|$ est la matrice des moments d'ordre deux de Q_l, $l = 1, 2$, alors en chassant les parenthèses dans (5) pour $\alpha = 0$, on trouve pour tous $a_1, \ldots, a_k$

$$\sum_{i,j=1}^{k} d_{ij}^{(1)} a_i a_j \leqslant \sum_{i,j=1}^{k} d_{ij}^{(2)} a_i a_j. \tag{6}$$

Dans le langage matriciel, cette relation s'écrit

$$d_1^2 \leqslant d_2^2. \tag{7}$$

Ceci exprime que la matrice $d_2^2 - d_1^1$ est semi-définie positive.

Donc, la dispersion quadratique moyenne de Q_1 autour de 0 est inférieure à celle de Q_2 si et seulement si les matrices des moments d'ordre deux vérifient les inégalités (6) et (7).

Dans le cas vectoriel la règle de préférence des estimateurs peut être formulée comme suit :

Approche de la moyenne quadratique : un estimateur θ_1^* est meilleur qu'un estimateur θ_2^* si sa dispersion quadratique moyenne autour de θ est strictement inférieure à celle de θ_2^*.

Si d_l^2 est la matrice des moments d'ordre deux de $\theta_l^* - \theta$, cette assertion se traduit par l'inégalité $d_1^2 < d_2^2$.

Approche asymptotique : un estimateur θ_1^* est meilleur qu'un estimateur θ_2^* si la dispersion quadratique moyenne autour de 0 de la distribution limite de $(\theta_1^* - \theta)\sqrt{n}$ est strictement inférieure à celle de la distribution limite de $(\theta_2^* - \theta)\sqrt{n}$.

Autrement dit, si $(\theta_l^* - \theta)\sqrt{n} \in \Phi_{0,\sigma_l^2}$, cette assertion exprime que $\sigma_1^2 < \sigma_2^2$.

On démontre que si θ_1^* et θ_2^* sont deux estimateurs asymptotiquement normaux et θ_1^* est meilleur que θ_2^*, alors

$$\lim_{n\to\infty} \mathbf{P}((\theta_1^* - \theta)\sqrt{n} \in B) > \lim_{n\to\infty} \mathbf{P}((\theta_2^* - \theta)\sqrt{n} \in B) \tag{8}$$

pour tout ellipsoïde *) central B.

Nous voyons que dans les deux approches, la comparaison des estimateurs revient à établir des inégalités entre matrices des moments d'ordre deux. La différence, c'est que dans la première approche les moments ne sont pas obligatoirement centrés.

---

*) Pour simplifier on conviendra d'appeler *ellipsoïde* dans R^k le domaine $\sum_{i,j=1}^{k} d_{ij}x_ix_j \leqslant$ $\leqslant c$ et *ellipse* la surface $\sum_{i,j=1}^{k} d_{ij}x_ix_j = c$.

Etablissons maintenant quelques relations équivalentes à (6) et (7). Posons

$$v(\theta^*) = \mathbf{E}(\theta^* - \theta)V(\theta^* - \theta)^T$$

et désignons par $\mathfrak{B}_+$ l'ensemble de toutes les matrices semi-définies positives $V = \|v_{ij}\|$. Si $\|d_{ij}\|$ est la matrice des moments d'ordre deux de $\theta^* - \theta$, il est alors évident que $v(\theta^*) = \sum v_{ij}d_{ij}$.

LEMME 1. *Pour que $d_1^2 \leqslant d_2^2$, il est nécessaire et suffisant que $v(\theta_1^*) \leqslant \leqslant v(\theta_2^*)$ pour toute $V \in \mathfrak{B}_+$.*

DÉMONSTRATION. La condition suffisante est évidente, puisque $V_a = = \|a_ia_j\| \in \mathfrak{B}_+$ et pour une telle matrice

$$v_a(\theta_l^*) = \mathbf{E}(\theta_l^* - \theta)V_a(\theta_l^* - \theta)^T = \sum a_ia_jd_{ij}^{(l)}$$

(cf. (6)).

Pour prouver la condition nécessaire, on remarquera que l'ordre partiel induit par les inégalités (5) est invariant par une rotation des axes de coordonnées. Plus exactement, si C est la matrice associée à une transformation orthogonale et si θ_1^* est meilleur que θ_2^*, alors θ_1^*C est meilleur que θ_2^*C. Ceci résulte des égalités

$$(\theta_l^*C - \theta C, a) = ((\theta_l^* - \theta)C, a) = (\theta_l^* - \theta, aC^T)$$

et de la définition 2.

Supposons maintenant que $d_1^2 < d_2^2$, c'est-à-dire que

$$\sum d_{ij}^{(1)}a_ia_j < \sum d_{ij}^{(2)}a_ia_j. \tag{9}$$

Ceci exprime que $v(\theta_1^*) < v(\theta_2^*)$ pour les matrices V de la forme $V_a = \|a_ia_j\|$ et donc pour les matrices diagonales $V_{\mathrm{diag}} \in \mathfrak{B}_+$, puisque ces dernières se représentent par la somme de k matrices de la forme V_a. Supposons maintenant que V est une matrice arbitraire de $\mathfrak{B}_+$ et C une transformation orthogonale telle que $C^TVC = V_{\mathrm{diag}}$. Alors

$$v(\theta_1^*) = \mathbf{E}(\theta_1^* - \theta)V(\theta_1^* - \theta)^T = \mathbf{E}(\theta_1^* - \theta)CV_{\mathrm{diag}}C^T(\theta_1^* - \theta)^T.$$

Des deux remarques ci-dessus et de (9) il s'ensuit que le second membre de cette égalité est strictement inférieur à

$$\mathbf{E}(\theta_2^* - \theta)CV_{\mathrm{diag}}C^T(\theta_2^* - \theta)^T = \mathbf{E}(\theta_2^* - \theta)V(\theta_2^* - \theta)^T = v(\theta_2^*). \blacktriangleleft$$

Il existe un autre procédé de comparaison des dispersions quadratiques moyennes (cf. [19]) qui implique néanmoins que les distributions $\mathbf{Q}_1$ et $\mathbf{Q}_2$ ne soient pas dégénérées dans R^k et admettent des moyennes nulles. Dans ce cas les matrices des moments centrés d'ordre deux d_l^2 seront définies positives et admettront les matrices inverses $A_l = (d_l^2)^{-1}$.

Soit d^2 la matrice des moments d'ordre deux de la distribution **Q** et soit $A = (d^2)^{-1}$.

DÉFINITION 3. On appelle *ellipsoïde de dispersion de la distribution* **Q** l'unique ellipsoïde

$$tAt^T \leqslant k + 2$$

sur lequel sont confondus les moments d'ordre un et d'ordre deux de **Q** et de **U**, **U** étant une distribution uniforme sur cet ellipsoïde (c'est-à-dire une distribution dans R^k de densité constante à l'intérieur de l'ellipsoïde et nulle en dehors) (cf. [19]).

LEMME 2. *Supposons que les matrices d_l^2, $l = 1, 2$, ne sont pas dégénérées. La dispersion quadratique moyenne de* $\mathbf{Q}_1$ *autour de* 0 *est inférieure à celle de* $\mathbf{Q}_2$ *si et seulement si l'ellipsoïde de dispersion de* $\mathbf{Q}_1$ *est contenu dans celui de* $\mathbf{Q}_2$.

DÉMONSTRATION. Supposons que l'ellipse $tA_1t^T = 1$ est contenue dans l'ellipse $tA_2t^T = 1$. On sait qu'il existe une application linéaire non dégénérée $t = uL$ qui envoie l'ellipse $tA_1t^T = 1$ dans la sphère unité S_1 et l'ellipse $tA_2t^T = 1$ dans une ellipse S_2 dont les axes principaux sont de même direction que les axes de coordonnées. Ceci exprime que $\tilde{A}_1 \equiv LA_1L^T = E$ (E est la matrice unité), $\tilde{A}_2 \equiv LA_2L^T = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_k^2)$, $0 < \lambda_j^2 \leqslant 1$, $j = 1, \ldots, k$. Puisque $\tilde{A}_1^{-1} = E$ et $\tilde{A}_2^{-1} = \operatorname{diag}(\lambda_1^{-2}, \ldots, \lambda_k^{-2})$, l'ellipse $t\tilde{A}_2^{-1}t^T = 1$ sera l'inverse de l'ellipse S_2 par rapport à la sphère unité S_1, et par suite, elle sera contenue dans S_1. Comme $\tilde{A}_2^{-1} = (L^T)^{-1}A_2L^{-1}$, en effectuant la transformation inverse $u = tL^T$, on trouve que l'ellipse $tA_1^{-1}t^T = td_1^2t = 1$ est située à l'extérieur de $tA_2^{-1}t^T = td_2^2t^T = 1$. Il est évident que cette relation est valable pour les ellipses $td_1^2t^T = c$ et $td_2^2t^T = c$. Or cela signifie que l'égalité $td_1^2t^T = c$ entraîne $td_1^2t^T = c \leqslant td_2^2t^T$. La réciproque se démontre de façon analogue. ◄

Il est important de signaler que contrairement au cas scalaire, la comparaison des dispersions des distributions à l'aide des matrices des moments d'ordre deux n'induit qu'une relation d'ordre partiel sur l'ensemble des distributions. Par exemple, des deux matrices $d_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ et $d_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ aucune n'est ni pire ni meilleure que l'autre, puisque le vecteur $a = (1, 0)$ vérifie l'inégalité (6) tandis que le vecteur $a = (0, 1)$, l'inégalité contraire. Ceci est un grave inconvénient de l'ordre introduit, bien que son adéquation ne fasse pas de doute.

On peut totalement ordonner un ensemble d'estimateurs (ou un ensemble de distributions) si l'on y compare par exemple $\mathbf{E}|\theta^* - \theta|^2$, où $|\cdot|$ est

une norme euclidienne sur R^k, de sorte que

$$\mathbf{E}|\theta^* - \theta|^2 = \mathbf{E} \sum_{i=1}^{k} (\theta_i^* - \theta_i)^2. \tag{10}$$

Cette façon d'ordonner l'ensemble est déjà contestable, puisque la précision s'estime différemment selon les cas et les directions. Pour prendre cette circonstance en compte, on peut, en guise de généralisation, considérer le critère de précision

$$v(\theta^*) = \mathbf{E}(\theta^* - \theta)V(\theta^* - \theta)^T,$$

où V est une matrice semi-définie positive (le cas (10) correspond à $V = E$).

Du lemme 1 il s'ensuit que si la dispersion de θ_1^* autour de θ est strictement inférieure à celle de θ_2^*, alors $v(\theta_1^*) < v(\theta_2^*)$. La réciproque n'est généralement pas vraie : la réalisation de l'inégalité $v(\theta_1^*) < v(\theta_2^*)$ pour une matrice V donnée quelconque (l'ordre total proposé ci-dessus est basé sur une matrice fixe) ne signifie encore pas que la dispersion de θ_1^* autour de θ soit strictement inférieure à celle de θ_2^*.

Passons maintenant à l'étude d'un cas *paramétrique* important impliquant l'estimation de paramètres inconnus de distributions appartenant à des familles paramétriques.

§ 8. Comparaison des estimateurs dans le cas paramétrique. Estimateurs efficaces

Dans le paragraphe précédent nous avons dégagé deux approches (de la moyenne quadratique et asymptotique) de comparaison de la qualité des estimateurs. Introduisons quelques notions liées à ces approches dans le cas où la distribution de l'échantillon X appartient à une famille paramétrique $\mathscr{P} = \{\mathbf{P}_\theta\}$. Les symboles $\mathbf{E}_\theta$ et $\mathbf{V}_\theta$ désigneront comme toujours l'espérance mathématique et la variance de la distribution $\mathbf{P}_\theta$.

1. Cas scalaire. On rappelle qu'aux termes de l'approche de la moyenne quadratique on dira que θ_1^* est meilleur que θ_2^* si

$$d_1^2(\theta) = \mathbf{E}_\theta(\theta_1^* - \theta)^2 < \mathbf{E}_\theta(\theta_2^* - \theta)^2 = d_2^2(\theta). \tag{1}$$

Mais dans le cas paramétrique, $d_l^2(\theta)$, $l = 1, 2$, sont des fonctions de θ et l'on dira que « θ_1^* est meilleur que θ_2^* *au point* θ » si $d_1(\theta) < d_2(\theta)$.

La situation est identique lorsque dans le cadre de l'approche asymptotique on compare les estimateurs asymptotiquement normaux au moyen de leurs distributions limites. Un estimateur θ_1^* est meilleur qu'un estimateur

θ_2^* au point θ^* si pour les variances σ_l, $l = 1, 2$, intervenant dans les relations

$$(\theta_l^* - \theta)\sqrt{n} \in \Phi_{0,\, \sigma_l^2(\theta)},\ l = 1, 2, \tag{2}$$

on a $\sigma_1(\theta) < \sigma_2(\theta)$ *).

Donc, dans les deux cas, la comparaison des estimateurs se ramène à la *comparaison de fonctions* : par exemple les fonctions $d_l(\theta)$, $\theta \in \Theta$. Cet ensemble d'estimateurs peut être muni d'un ordre partiel de la manière suivante.

RÈGLE 1. *Un estimateur θ_1^* est meilleur que θ_2^* si $d_1(\theta) \leqslant d_2(\theta)$ (ou respectivement $\sigma_1(\theta) \leqslant \sigma_2(\theta)$) pour tous les $\theta \in \Theta$ et si $d_1(\theta) < d_2(\theta)$ pour au moins un θ.*

S'il existe un estimateur θ_1^* meilleur que θ^*, on dira que θ^* est un *estimateur inadmissible*.

Arrêtons-nous tout d'abord sur l'approche de la moyenne quadratique dans le cas scalaire et étudions les possibilités de comparaison des estimateurs. Signalons d'entrée qu'il n'existe pas de meilleur estimateur au sens de la définition mentionnée. Autrement dit, il n'existe pas d'estimateur θ^* tel que pour tout autre estimateur θ_1^* l'on ait $d(\theta) \leqslant d_1(\theta)$, où $d_1(\theta)$ est définie dans (1) et $d(\theta)$ correspond à θ^*.

En effet, si $\theta_1^* = \theta_1 = \text{const} \in \Theta$, alors $d_1^2(\theta) = \mathbf{E}_\theta(\theta_1^* - \theta)^2 = 0$ pour $\theta = \theta_1$, et pour le meilleur estimateur θ^* (s'il existe) on aura $d^2(\theta_1) = \mathbf{E}_{\theta_1}(\theta^* - \theta_1)^2 = 0$. Comme θ_1 est arbitraire, il vient $d^2(\theta) \equiv 0$. Or ceci n'est possible que dans le cas « dégénéré » où les observations définissent de façon unique la valeur du paramètre θ. Par exemple, lorsque $X \in \mathbf{I}_\theta$ ou $X \in \mathbf{U}_{\theta,\, \theta+1}$ et $\Theta = \{1, 2, \ldots\}$.

Donc, l'enveloppe inférieure des fonctions $d^2(\theta)$ est nulle, mais cette fonction n'est réalisée pour aucun estimateur θ^* dans les cas « non dégénérés ».

Ce problème peut être rendu plus consistant si l'on cherche les meilleurs estimateurs θ^* dans des *sous-classes* d'estimateurs choisies de façon assez raisonnable. Une méthode de détermination de ces sous-classes consiste à fixer le biais $b(\theta)$.

DÉFINITION 1. On dit qu'un estimateur $\theta_0^* \in K$ est *efficace dans la classe* K si $\mathbf{E}_\theta(\theta_0^* - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta)^2$ quels que soient $\theta^* \in K$ et $\theta \in \Theta$.

La classe K_0 des estimateurs sans biais (c'est-à-dire tels que $b(\theta) \equiv 0$) joue un rôle particulier.

---

*) Nous avons déjà signalé que dans un grand nombre de cas $d_l^2(\theta) = n^{-1}\sigma_l^2 + o(n^{-1})$. Mais ceci ne résulte pas de la définition des nombres $d_l^2(\theta)$ et $\sigma_l^2(\theta)$.

Les estimateurs efficaces de la classe $K_0 = \{\theta^* : \mathbf{E}_\theta\theta^* = \theta\}$ des estimateurs sans biais sont dits simplement *efficaces*. En sorte que les estimateurs efficaces sont des estimateurs sans biais, de variance minimale.

La propriété d'être sans biais est, comme on l'a déjà signalé, une propriété qui est souhaitable en soi dans la mesure où elle exprime l'absence de toute erreur systématique.

Pour s'assurer de l'existence l'estimateurs de biais $b(\theta)$ donné (en particulier d'estimateurs sans biais) il faut résoudre une équation intégrale par rapport à $g(x)$:

$$\int g(x)\mathbf{P}_\theta(X \in dx) = \theta + b(\theta), \tag{3}$$

où $g(X) = \theta^*$; le premier membre de cette équation est $\mathbf{E}_\theta\theta^*$.

Si la condition (A_μ) est réalisée et si la fonction $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$ est une fonction de vraisemblance, l'équation (3) devient

$$\int g(x)f_\theta(x)\mu^n(dx) = \theta + b(\theta). \tag{4}$$

Signalons que l'équation (4) n'admet pas toujours une solution pour $b(\theta)$ donné et qu'en particulier *les estimateurs sans biais du paramètre θ n'existent pas pour toutes les familles* $\{\mathbf{P}_\theta\}$. Considérons par exemple un schéma de Bernoulli de paramètre inconnu p (la probabilité de l'issue $\{x_1 = 1\}$) et soit à estimer un paramètre $\theta = \varphi(p)$, où φ est une fonction donnée. L'équation (4) devient alors

$$\sum_x g(x)f_\theta(x) = \theta$$

ou ce qui est équivalent

$$\sum_{k=0}^{n} G(k)p^k(1 - p)^{n-k} = \varphi(p), \tag{5}$$

où $G(k) = \sum_{x \in A_k} g(x)$, A_k est l'ensemble des points x dont k coordonnées sont égales à 1. Or le premier membre de (5) est un polynôme en p de degré n, donc l'équation (5) admet une solution si seulement $\varphi(p)$ est un polynôme de degré $\leqslant n$.

Considérons maintenant la classe K_b des estimateurs de biais $b(\theta)$ fixé et supposons qu'il existe un estimateur efficace dans K_b.

THÉORÈME 1. *Il existe un seul estimateur efficace de K_b aux valeurs près sur un ensemble $A \subset \mathscr{X}^n$ tel que $\mathbf{P}_\theta(A) = 0$ pour tous les $\theta \in \Theta$.*

DÉMONSTRATION. Soient θ_0^* et θ_1^* deux estimateurs efficaces dans K_b.

Posons

$$D = \mathbf{V}_\theta \theta_l^*, \quad \Delta_l = \theta_l^* - \theta, \quad \theta^* = \frac{\theta_0^* + \theta_1^*}{2}, \quad l = 0, 1.$$

Puisque

$$\left(\frac{\Delta_0 + \Delta_1}{2}\right)^2 + \left(\frac{\Delta_0 - \Delta_1}{2}\right)^2 = \frac{\Delta_0^2 + \Delta_1^2}{2},$$
$$\frac{\Delta_0 + \Delta_1}{2} = \theta^* - \theta, \; \Delta_0 - \Delta_1 = \theta_0^* - \theta_1^*, \tag{6}$$

il vient

$$\mathbf{E}_\theta(\theta^* - \theta)^2 + \frac{1}{4}\mathbf{E}_\theta(\theta_0^* - \theta_1^*)^2 = D + b^2(\theta). \tag{7}$$

Or $\theta^* \in K_b$. Donc, $\mathbf{E}_\theta(\theta^* - \theta)^2 \geqslant D + b^2(\theta)$. De (7) il s'ensuit alors que

$$\mathbf{E}_\theta(\theta_0^* - \theta_1^*)^2 \leqslant 0,$$

$\theta_1^* = \theta_0^*$ presque partout *). ◄

Restons encore dans le cadre de l'approche de la moyenne quadratique et introduisons la

DÉFINITION 2. On dit qu'un estimateur $\theta_1^* \in K$ est *asymptotiquement efficace dans* K si

$$\limsup_{n\to\infty} \frac{\mathbf{E}_\theta(\theta_1^* - \theta)^2}{\mathbf{E}_\theta(\theta^* - \theta)^2} \leqslant 1, \tag{8}$$

quels que soient $\theta^* \in K$ et $\theta \in \Theta$.

Passons maintenant à l'approche asymptotique à laquelle la définition 2 est aussi étroitement liée. Le problème consiste comme précédemment à comparer des fonctions $\sigma(\theta)$ caractérisant la distribution normale limite, mais la situation est dans l'ensemble légèrement simplifiée. D'abord parce que la comparaison est effectuée dans la classe des estimateurs asymptotiquement normaux, qui sera désignée ultérieurement par K_Φ. On peut restreindre cette classe sans l'appauvrir. Plus exactement, on étudiera la classe

---

*) On a la proposition suivante qui généralise dans un certain sens le théorème 1. *Si* θ_0^* *est un estimateur efficace de* K_b *et* θ^* *un estimateur de* K_b *tel que* $h = \mathbf{V}_\theta\theta_0^*/\mathbf{V}_\theta\theta^* \leqslant 1$, *alors le coefficient de corrélation* $\varrho(\theta_0^*, \theta^*)$ *entre* θ_0^* *et* θ^* *est égal à* $\sqrt{h}$.

Le lecteur pourra s'exercer à prouver ce fait en s'assurant que, si $\varrho(\theta_0^*, \theta^*) \neq \sqrt{h}$ et si α est convenablement choisi, l'estimateur

$$\theta_1^* = (1 - \alpha)\theta_0^* + \alpha\theta^* \in K_b$$

vérifiera l'inégalité $\mathbf{V}_\theta\theta_1^* < \mathbf{V}_\theta\theta_0^*$ qui contredit l'efficacité de θ_0^*.

$K_{\Phi,2} \subset K_\Phi$ des estimateurs asymptotiquement normaux θ^* pour lesquels la convergence

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\,\sigma^2(\theta)}$$

a lieu en même temps que celle des moments d'ordre un et deux

$$\mathbf{E}_\theta(\theta^* - \theta)\sqrt{n} \to 0, \quad \mathbf{E}_\theta(\theta^* - \theta)^2 n \to \sigma^2(\theta). \tag{9}$$

A noter que la première de ces relations se déduit sans peine de la deuxième à l'aide du théorème de continuité pour les moments (§ 1.5).

La restriction de K_Φ à $K_{\Phi,2}$ appauvrit peu la première classe pour les deux raisons suivantes. Premièrement, les estimateurs asymptotiquement normaux violant (9) ne se rencontrent pratiquement pas (nous avons vu qu'ils impliquaient des constructions artificielles). Deuxièmement, d'après le lemme de Fatou, pour $\theta^* \in K_\Phi$, on a

$$\lim_{n\to\infty} \inf \mathbf{E}_\theta n(\theta^* - \theta)^2 \geqslant \sigma^2(\theta)$$

(nous avons affaire à des intégrales de fonctions positives), de sorte que $\mathbf{E}_\theta n(\theta^* - \theta)^2$ peut être seulement supérieure à $\sigma^2(\theta)$ pour les grands n. Or il est peu probable que les estimateurs doués de cette propriété puissent concurrencer les estimateurs vérifiant (9).

Donc, dans l'approche asymptotique, pour classe des estimateurs asymptotiquement normaux dans laquelle a lieu la comparaison, on peut prendre la classe $K_{\Phi,2}$ pour la commodité.

Soit K une classe d'estimateurs telle que $K \subset K_{\Phi,2}$. La définition suivante est équivalente à la définition 2.

DÉFINITION 3. On dit qu'un estimateur $\theta_1^* \in K$ est *asymptotiquement efficace dans* K si quels que soient $\theta^* \in K$ et $\theta \in \Theta$

$$\sigma_1^2(\theta) \leqslant \sigma^2(\theta) \tag{10}$$

où $\sigma^2(\theta)$ et $\sigma_1^2(\theta)$ sont les paramètres de dispersion de θ^* et θ_1^* respectivement.

L'équivalence de ces définitions résulte du fait que pour $\theta^* \in K_{\Phi,2}$, on a

$$\mathbf{E}_\theta(\theta^* - \theta)^2 = \frac{\sigma^2(\theta)}{n}(1 + r_n(\theta)), \quad r_n(\theta) \to 0 \quad \text{pour} \quad n \to \infty.$$

Dans ce cas, la relation (8) qui exprime que

$$\mathbf{E}_\theta(\theta_1^* - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta)^2(1 + r_n'(\theta)), \quad r_n'(\theta) \to 0,$$

pour tout $\theta^* \in K$, est visiblement équivalente à l'inégalité (10). ◄

La simplification évoquée dans l'approche asymptotique consiste en la comparaison des seules variances des lois limites. Le biais n'importe plus puisque dans la classe $K_{\Phi,2}$, en vertu de (9), est réalisée la relation $b(\theta) = o(1/\sqrt{n})$ qui exprime que les estimateurs sont « pratiquement sans biais »

ou que le biais est « asymptotiquement négligeable » du point de vue des relations (2).

Par analogie au théorème 1, on a le

THÉORÈME 2. *Soit $K \subset K_{\Phi,2}$. Si θ_1^* et θ_2^* sont deux estimateurs asymptotiquement efficaces de K tels que $\frac{1}{2}(\theta_1^* + \theta_2^*) \in K$, ils sont alors asymptotiquement confondus, c'est-à-dire que*

$$\sqrt{n}(\theta_1^* - \theta_2^*) \underset{P}{\rightarrow} 0, \quad \mathbf{E}_\theta[\sqrt{n}(\theta_1^* - \theta_2^*)]^2 \rightarrow 0.$$

DÉMONSTRATION. Il nous suffit de prouver la deuxième relation, puisque la première en est une conséquence. Soit

$$M_{l,n} = \mathbf{E}_\theta n(\theta_l^* - \theta)^2, \quad \Delta_l = \theta_l^* - \theta, \; \theta^* = \frac{\theta_1^* + \theta_2^*}{2}, \; l = 1, 2.$$

En vertu de (6), on obtient alors

$$\mathbf{E}_\theta n(\theta^* - \theta)^2 + \frac{1}{4}\mathbf{E}_\theta n(\theta_1^* - \theta_2^*)^2 = (M_{1,n} + M_{2,n})/2. \tag{11}$$

Or $\theta^* \in K$, donc en passant à la limite dans la dernière égalité, on trouve en vertu de l'efficacité asymptotique de θ_l^* que

$$\lim_{n\to\infty} \mathbf{E}_\theta n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \blacktriangleleft$$

Les considérations développées ci-dessus n'indiquaient qu'une des éventuelles méthodes d'acquisition des estimateurs (des estimateurs efficaces ici) qui sont meilleurs que les autres pour des critères naturels. Il existe toutefois d'autres méthodes (signalons qu'il nous faut comparer des éléments non ordonnés : des fonctions $d(\theta)$ ou $\sigma(\theta)$). Etant donné qu'il n'existe pas en général d'estimateurs dont $d(\theta)$ soit minimale pour chaque θ, on peut comparer par exemple les valeurs moyennes $\int d(t)q(t)dt$, où $q(t) \geqslant 0$, $\int q(t)dt = 1$, ou les valeurs maximales $\max_{\theta\in\Theta} d(\theta)$. Nous avons là deux façons d'ordonner l'ensemble de *tous* les estimateurs.

Le premier procédé sera appelé ultérieurement *bayésien*, le second, *minimax*. Les estimateurs optimaux bayésiens et minimax seront traités au § 11, les estimateurs efficaces, dans les paragraphes suivants.

Le problème du choix des estimateurs sera examiné avec plus de détails à la fin de cet ouvrage (cf. avant-propos).

2. **Cas vectoriel.** Considérons maintenant le cas où θ et θ^* sont des vecteurs de R^k. La comparaison des estimateurs est plus délicate ici. En effet, dans le cas vectoriel, nous avons été contraints d'introduire un ordre *partiel*

pour comparer les estimateurs à θ *fixe*. Pour comparer les estimateurs sur l'ensemble Θ *tout entier*, il nous faut, comme dans le cas scalaire, introduire un ordre partiel mais dans une « autre direction » (la comparaison est effectuée à l'aide de l'écart quadratique moyen qui est une fonction de deux variables : θ et le vecteur a sur lequel est projeté l'écart $\theta^* - \theta$).

Les meilleurs estimateurs dans ces « deux directions » font l'objet des définitions suivantes.

DÉFINITION 4. On dit qu'un estimateur θ_0^* est *efficace dans la classe K* si la dispersion quadratique moyenne de θ^* autour de θ est supérieure à celle de θ_0^* quels que soient $\theta^* \in K$ et $\theta \in \Theta$.

Cette définition équivaut à la suivante.

Un estimateur vectoriel θ_0^* de θ est efficace dans K si pour tout vecteur a l'estimateur $\alpha_0^* = (\theta_0^*, a)$ est un estimateur efficace du paramètre $\alpha = (\theta, a)$ dans la classe des estimateurs $\alpha^* = (\theta^*, a)$, $\theta^* \in K$, c'est-à-dire que pour tous les $\theta \in \Theta$, $a \in R^k$, $\theta^* \in K$

$$\mathbf{E}_\theta(\theta_0^* - \theta, a)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta, a)^2. \tag{12}$$

Nous avons vu que cette inégalité peut s'écrire sous la forme équivalente $d_0^2(\theta) \leqslant d^2(\theta)$ ou

$$\sum_{i,j} d_{ij}^{(0)}(\theta)a_i a_j \leqslant \sum_{i,j} d_{ij} a_i a_j$$

pour tous les $\theta \in \Theta$, $a \in R^k$, où $d^2(\theta) = \| d_{ij}(\theta) \|$ et $d_0^2(\theta) = \| d_{ij}^{(0)}(\theta) \|$ sont les matrices des moments d'ordre deux de $\theta^* - \theta$ et $\theta_0^* - \theta$ respectivement.

Les estimateurs efficaces de la classe K_0 des estimateurs sans biais sont tout simplement dits *efficaces*.

Etant donné que la définition (12) de l'efficacité est fondée sur l'utilisation du cas scalaire, le théorème 1 nous permet d'établir immédiatement qu'il existe un seul estimateur efficace dans la classe K_b des estimateurs de biais fixe $b(\theta) = \mathbf{E}\theta^* - \theta$.

La définition des estimateurs asymptotiquement efficaces dans le cas vectoriel est calquée sur les définitions 2, 3.

DÉFINITION 5. On dit qu'un estimateur vectoriel θ_1^* d'un paramètre θ est *asymptotiquement efficace dans K* si pour tout vecteur a, l'estimateur (θ_1^*, a) est un estimateur asymptotiquement efficace du paramètre scalaire $\alpha = (\theta, a)$ dans la classe des estimateurs $\alpha^* = (\theta^*, a)$, $\theta^* \in K$.

En d'autres termes (cf. § 7), la dispersion quadratique moyenne de la distribution limite de $(\theta_1^* - \theta)\sqrt{n}$ est minimale pour les estimateurs asymptotiquement efficaces. Ce qui exprime que pour tous $\theta^* \in K$, $a \in R^k$, $\theta \in \Theta$, on a $\sigma_1^2(\theta) \leqslant \sigma^2(\theta)$, ou

$$\sum_{i,j} \sigma_{ij}^{(1)}(\theta)a_i a_j \leqslant \sum_{i,j} \sigma_{ij}(\theta)a_i a_j,$$

où $\sigma^2(\theta) = \|\sigma_{ij}(\theta)\|$, $\sigma_1^2(\theta) = \|\sigma_{ij}^{(1)}(\theta)\|$ sont respectivement les matrices des moments d'ordre deux des distributions limites de $(\theta^* - \theta)\sqrt{n}$ et de $(\theta_1^* - \theta)\sqrt{n}$.

Du paragraphe précédent on déduit que dans le cas vectoriel l'ensemble des estimateurs à θ fixe peut être ordonné si la qualité d'un estimateur est mesurée (dans l'approche de la moyenne quadratique) par

$$v(\theta^*) = \mathsf{E}_0(\theta^* - \theta)V(\theta^* - \theta)^T \equiv v(\theta^*, \theta), \tag{13}$$

où V est une matrice semi-définie positive. On pourrait envisager une quantité analogue faisant intervenir la matrice des moments d'ordre deux de la loi normale limite dans le cas de l'approche asymptotique dans la classe $K_{\Phi,2}$.

En poussant plus loin dans cette voie, on peut ordonner totalement l'ensemble de tous les estimateurs pour l'ensemble Θ tout entier. Plus exactement, on peut comparer les moyennes

$$\int v(\theta^*, t)\, q(t)dt, \quad q(t) \geqslant 0, \quad \int q(t)dt = 1,$$

ou les valeurs $\max\limits_{t\in\Theta} v(\theta^*, t)$ de $v(\theta^*, \theta)$ définies dans (13).

Si le meilleur estimateur au sens de cette approche restera le meilleur pour toute matrice V semi-définie positive, c'est que, en vertu du lemme 7.1, il sera le meilleur au sens de l'ordre partiel défini au § 7 (autrement dit la moyenne de la dispersion quadratique moyenne sera minimale dans toute direction).

Pour construire des estimateurs optimaux au sens des définitions envisagées dans ce paragraphe, nous aurons besoin des notions et des propriétés des espérances mathématiques conditionnelles et des statistiques exhaustives.

§ 9. Espérances mathématiques conditionnelles

Dans ce paragraphe, on rappelle la définition et les principales propriétés des espérances mathématiques conditionnelles. Pour un exposé plus complet voir Annexe III ainsi que [11], [17], [24], [34], [53].

1. Définition de l'espérance mathématique conditionnelle. Soient ξ et η deux variables aléatoires définies sur un espace probabilisé $(\Omega, \mathfrak{F}, \mathbf{P})$.

L'espérance mathématique conditionnelle $\mathsf{E}(\xi | B)$ de la variable aléatoire ξ par rapport à l'événement B, $\mathbf{P}(B) > 0$, est définie par la relation

$$\mathsf{E}(\xi | B) = \frac{\mathsf{E}(\xi\,;B)}{\mathbf{P}(B)}, \tag{1}$$

où $\mathsf{E}(\xi\,;B) = \int_B \xi\, d\mathbf{P} = \mathsf{E}(\xi I_B)$, $I_B = I_B(\omega)$ est une variable aléatoire égale à l'indicateur de l'ensemble B.

Supposons que ξ et η sont indépendantes, $B = \{\eta = x\}$ et $\mathbf{P}(B) > 0$. Pour toute fonction mesurable $\varphi(x, y)$, on a alors en vertu de (1)

$$\mathbf{E}[\varphi(\xi, \eta)|\eta = x] = \frac{\mathbf{E}\varphi(\xi, \eta)I_{\{\eta = x\}}}{\mathbf{P}(\eta = x)} = \frac{\mathbf{E}\varphi(\xi, x)I_{\{\eta = x\}}}{\mathbf{P}(\eta = x)} = \mathbf{E}\varphi(\xi, x). \quad (2)$$

La validité de la dernière égalité découle de l'indépendance des variables aléatoires $\varphi(\xi, x)$ et $I_{\{\eta = n\}}$ en tant que fonctions de ξ et η respectivement, et par suite

$$\mathbf{E}\varphi(\xi, x)I_{\{\eta = x\}} = \mathbf{E}\varphi(\xi, x)\mathbf{E}I_{\{\eta = x\}} = \mathbf{E}\varphi(\xi, x)\mathbf{P}(\eta = x).$$

Les relations (2) montrent que la notion d'espérance mathématique conditionnelle garde son sens dans le cas aussi où la probabilité de la condition est nulle : en effet l'égalité

$$\mathbf{E}[\varphi(\xi, \eta)|\eta = x] = \mathbf{E}\varphi(\xi, x)$$

pour ξ et η indépendantes est naturelle en soi et n'est en aucune façon liée à l'hypothèse que $\mathbf{P}(\eta = x) > 0$.

Soit $\mathfrak{A}$ une sous-tribu de $\mathfrak{F}$. Définissons la notion d'espérance mathématique conditionnelle $\mathbf{E}(\xi|\mathfrak{A})$ d'une variable aléatoire ξ par rapport à $\mathfrak{A}$. Nous donnerons tout d'abord cette définition dans le cas « discret » sous une forme facilement généralisable.

Nous appellerons « discret » le cas où la tribu $\mathfrak{A}$ est engendrée par une suite au plus dénombrable d'événements disjoints $A_1, A_2, \ldots$; $\bigcup A_i = \Omega$, $\mathbf{P}(A_i) > 0$. On notera ce fait par le symbole $\mathfrak{A} = \sigma(A_1, A_2, \ldots)$ qui exprime que les éléments de $\mathfrak{A}$ sont toutes les réunions possibles des ensembles A_1, $A_2, \ldots$

Construisons une nouvelle variable aléatoire $\hat{\xi} = \hat{\xi}(\omega)$ à l'aide de ξ et du système d'événements $(A_1, A_2, \ldots)$ de la manière suivante :

$$\hat{\xi} = y_k \equiv \mathbf{E}(\xi|A_k) = \frac{\mathbf{E}(\xi\,;A_k)}{\mathbf{P}(A_k)} \quad \text{pour} \quad \omega \in A_k, \quad k = 1, 2, \ldots$$

En d'autres termes

$$\hat{\xi} = \sum_k \frac{\mathbf{E}(\xi\,;A_k)}{\mathbf{P}(A_k)} I_{A_k}$$

où I_A est l'indicateur de l'ensemble A.

DÉFINITION 1. La variable aléatoire $\hat{\xi}$ s'appelle *espérance mathématique conditionnelle de* ξ *par rapport à la tribu* $\mathfrak{A}$ et se note $\mathbf{E}(\xi\,|\,\mathfrak{A})$.

Ainsi, l'espérance mathématique conditionnelle $\mathbf{E}(\xi\,|\,\mathfrak{A})$ est une *variable aléatoire* contrairement aux espérances mathématiques ordinaires. Dans notre cas, elle est constante sur les ensembles A_k et égale à la moyenne de ξ sur A_k. Si ξ et $\mathfrak{A}$ sont indépendantes (autrement dit, si $\mathbf{P}(\xi \in B\,;A_k) =$ $= \mathbf{P}(\xi \in B)\mathbf{P}(A_k)$), il est alors évident que $\mathbf{E}(\xi\,;A_k) = \mathbf{E}\xi\mathbf{P}(A_k)$ et $\hat{\xi} = \mathbf{E}\xi$.

Si $\mathfrak{A} = \mathfrak{F}$, alors $\mathfrak{F}$ est « discrète » aussi, ξ est constante sur les ensembles A_k, et donc $\hat{\xi} = \xi$.

Signalons les deux propriétés fondamentales suivantes de l'espérance mathématique conditionnelle :

1) *$\hat{\xi}$ est mesurable par rapport à $\mathfrak{A}$.*

2) *Pour tout événement $A \in \mathfrak{A}$, on a*

$$\mathbf{E}(\hat{\xi}\,;A) = \mathbf{E}(\xi\,;A).$$

La première propriété est évidente. La deuxième résulte du fait que tout événement $A \in \mathfrak{A}$ peut être représenté sous la forme $A = \bigcup_k A_{j_k}$, donc

$$\mathbf{E}(\hat{\xi}\,;A) = \sum_k \mathbf{E}(\hat{\xi}\,;A_{j_k}) = \sum_k y_{j_k}\mathbf{P}(A_{j_k}) = \sum_k \mathbf{E}(\xi\,;A_{j_k}) = \mathbf{E}(\xi\,;A).$$

Cette propriété est suffisamment claire : la moyennisation de ξ sur l'ensemble A nous fournit le même résultat que la moyennisation de la quantité $\hat{\xi}$ déjà moyennisée sur A_{j_k}.

LEMME 1. *Les propriétés* 1) *et* 2) *définissent de façon unique l'espérance mathématique conditionnelle et sont équivalentes à la définition* 1.

DÉMONSTRATION. Nous avons déjà prouvé que les propriétés 1) et 2) découlaient de la définition 1. Supposons maintenant que sont réalisées les conditions 1 et 2. La mesurabilité de $\hat{\xi}$ par rapport à $\mathfrak{A}$ exprime que $\hat{\xi}$ est constante sur les ensembles A_k. Désignons par y_k la valeur de $\hat{\xi}$ sur A_k. Comme $A_k \in \mathfrak{A}$, il s'ensuit de la propriété 2 que

$$\mathbf{E}(\hat{\xi}\,;A_k) = y_k\mathbf{P}(A_k) = \mathbf{E}(\xi\,;A_k),$$

donc, pour $\omega \in A_k$,

$$\hat{\xi} = y_k = \frac{\mathbf{E}(\xi\,;A_k)}{\mathbf{P}(A_k)}. \blacktriangleleft$$

Nous pouvons donner maintenant une définition générale de l'espérance mathématique conditionnelle.

DÉFINITION 2. Soient ξ une variable aléatoire sur un espace probabilisé $(\Omega, \mathfrak{F}, \mathbf{P})$ et $\mathfrak{A}$ une sous-algèbre de $\mathfrak{F}$. On appelle *espérance mathématique conditionnelle* $\mathbf{E}(\xi \mid \mathfrak{A})$ *de ξ par rapport à $\mathfrak{A}$* la variable aléatoire $\hat{\xi}$ dotée des deux propriétés suivantes :

1) *$\hat{\xi}$ est mesurable par rapport à $\mathfrak{A}$.*

2) $\mathbf{E}(\hat{\xi}\,;A) = \mathbf{E}(\xi\,;A)$ *pour tout $A \in \mathfrak{A}$.*

Dans cette définition, la variable ξ peut être aussi bien scalaire que vectorielle.

Deux questions viennent immédiatement à l'esprit : la variable $\hat{\xi}$ existe-t-elle et est-elle unique ? Dans le cas « discret » nous avons répondu par l'affirmative à ces questions. Dans le cas général, on a le

THÉORÈME 1. *Si* $\mathbf{E}\,|\,\xi\,|$ *est finie, la fonction* $\hat{\xi} = \mathbf{E}(\xi\,|\,\mathfrak{A})$ *de la définition 2 existe toujours et est unique presque partout, sauf peut-être sur un ensemble de probabilité nulle.*

DÉMONSTRATION. Supposons tout d'abord que ξ est scalaire, $\xi \geqslant 0$. La fonction d'ensemble

$$\mathbf{Q}(A) = \int_A \xi d\mathbf{P} = \mathbf{E}(\xi\,;A), \quad A \in \mathfrak{A},$$

sera alors une mesure sur $(\Omega, \mathfrak{A})$ absolument continue par rapport à $\mathbf{P}$, puisque $\mathbf{P}(A) = 0$ entraîne $\mathbf{Q}(A) = 0$. Donc, d'après le théorème de Radon-Nikodym ([11], Annexe 3), il existe une fonction $\hat{\xi} = \mathbf{E}(\xi\,|\,\mathfrak{A})$ $\mathfrak{A}$-mesurable, unique aux valeurs près sur un ensemble de mesure nulle, telle que

$$\mathbf{Q}(A) = \int_A \hat{\xi} d\mathbf{P}.$$

Dans le cas général, posons $\xi = \xi^+ - \xi^-$, $\xi^+ = \max(0, \xi) \geqslant 0$, $\xi^- = \max(0, -\xi) \geqslant 0$,

$$\hat{\xi} = \hat{\xi}^+ - \hat{\xi}^-,$$

où $\hat{\xi}^\pm$ est l'espérance mathématique conditionnelle de $\xi^\pm$. Ceci prouve l'existence de l'espérance mathématique conditionnelle, puisque $\hat{\xi}$ satisfera les conditions 1), 2) de la définition 2. D'où l'unicité de $\hat{\xi}$, puisque, le cas échéant, $\hat{\xi}^+$ et $\hat{\xi}^-$ ne seraient pas uniques. La démonstration pour les ξ vectorielles se ramène au cas scalaire, car les propriétés 1) et 2) seront satisfaites par les coordonnées de $\hat{\xi}$ dont l'existence et l'unicité ont été prouvées. ◄

L'idée de la démonstration réalisée est assez claire : d'après la condition 2 la quantité $\mathbf{E}(\hat{\xi}\,;A) = \int_A \hat{\xi} d\mathbf{P}$ est définie pour tout $A \in \mathfrak{A}$, c'est-à-dire que sont données les valeurs des intégrales de $\hat{\xi}$ sur tous les ensembles $A \in \mathfrak{A}$. Il est évident que ceci doit définir une fonction $\mathfrak{A}$- mesurable $\hat{\xi}$ qui est unique aux valeurs près sur un ensemble de mesure nulle.

Le sens de $\mathbf{E}(\xi | \mathfrak{A})$ reste le même : c'est *grosso modo* la moyenne de ξ sur les éléments « non divisibles » de $\mathfrak{A}$.

Si $\mathfrak{A} = \mathfrak{F}$, il est évident que $\hat{\xi} = \xi$ satisfait les conditions 1), 2), et par suite $\mathbf{E}(\xi | \mathfrak{F}) = \xi$.

DÉFINITION 3. Soient ξ et η des variables aléatoires sur $(\Omega, \mathfrak{F}, \mathbf{P})$, $\mathfrak{A} = \sigma(\eta)$ la tribu engendrée par η. Alors la quantité $\mathbf{E}(\xi | \mathfrak{A})$ s'appelle aussi *espérance mathématique conditionnelle de* ξ *par rapport à* η.

Pour simplifier l'écriture on écrira parfois $\mathsf{E}(\xi|\eta)$ au lieu de $\mathsf{E}(\xi|\sigma(\eta))$. Ceci n'entraînera aucune confusion.

Vu que $\mathsf{E}(\xi \mid \eta)$ est par définition une variable aléatoire $\sigma(\eta)$-mesurable, cela signifie (cf. [11]) qu'il existe une fonction mesurable $g(x)$ telle que

$$\mathsf{E}(\xi|\eta) = g(\eta). \tag{3}$$

Par analogie au cas discret, nous pouvons interpréter la quantité $g(x)$ comme la moyenne de ξ sachant que $\{\eta = x\}$. (On rappelle que dans le cas discret $g(x) = \mathsf{E}(\xi|\eta = x)$.)

DÉFINITION 4. Si $\xi = I_C$ est l'indicateur d'un ensemble $C \in \mathfrak{F}$, on dira que $\mathsf{E}(I_C|\mathfrak{A})$ est la *probabilité conditionnelle* $\mathbf{P}(C|\mathfrak{A})$ de l'événement C par rapport à $\mathfrak{A}$. Si $\mathfrak{A} = \sigma(\eta)$, on parlera de la probabilité conditionnelle $\mathbf{P}(C|\eta)$ de l'événement C par rapport à η.

2. Propriétés de l'espérance mathématique conditionnelle.

1) *L'espérance mathématique conditionnelle satisfait les propriétés ordinaires des espérances mathématiques* (cf. [11]) à la seule différence qu'elles sont réalisées presque sûrement :

1a) $\mathsf{E}(c\xi \mid \mathfrak{A}) = c\mathsf{E}(\xi \mid \mathfrak{A})$, *où* $c =$ const,

1b) $\mathsf{E}(\xi_1 + \xi_2 \mid \mathfrak{A}) = \mathsf{E}(\xi_1 \mid \mathfrak{A}) + \mathsf{E}(\xi_2 \mid \mathfrak{A})$,

1c) *si* $\xi_1 \leqslant \xi_2$ *presque sûrement, alors* $\mathsf{E}(\xi_1 \mid \mathfrak{A}) \leqslant \mathsf{E}(\xi_2 \mid \mathfrak{A})$.

2) *On a l'inégalité du genre inégalité de Tchébychev* : si $\xi \geqslant 0$ est réelle, pour tout $x > 0$ on a

$$\mathbf{P}(\xi \geqslant x \mid \mathfrak{A}) \leqslant \frac{\mathsf{E}(\xi \mid \mathfrak{A})}{x}.$$

Comme dans le n° 1 cette relation est réalisée presque sûrement. Cette convention sera valable pour toutes les relations mettant en jeu des espérances mathématiques conditionnelles.

3) *Si les tribus* $\mathfrak{A}$ *et* $\sigma(\xi)$ *sont indépendantes,* $\mathsf{E}(\xi \mid \mathfrak{A}) = \mathsf{E}\xi$.

De là il s'ensuit en particulier que si ξ et η sont indépendantes, on a $\mathsf{E}(\xi \mid \eta) = \mathsf{E}\xi$. Si la tribu $\mathfrak{A}$ est triviale, on obtient de toute évidence $\mathsf{E}(\xi \mid \mathfrak{A}) = = \mathsf{E}\xi$.

4) *Les espérances mathématiques conditionnelles vérifient les théorèmes de convergence pour les espérances mathématiques ordinaires.* Le théorème de convergence monotone, par exemple, s'énonce : si $\xi_n \uparrow \xi$, $\xi_n \geqslant 0$, alors $\mathsf{E}(\xi_n \mid \mathfrak{A}) \uparrow \mathsf{E}(\xi \mid \mathfrak{A})$ p.s.

5) *Si* η *est scalaire et* $\mathfrak{A}$*-mesurable,* $\mathsf{E}\,|\,\xi\,| < \infty$ *et* $\mathsf{E}\,|\,\xi\eta\,| < \infty$, *alors*

$$\mathsf{E}(\eta\xi \mid \mathfrak{A}) = \eta\mathsf{E}(\xi \mid \mathfrak{A}).$$

En d'autres termes, les variables aléatoires $\mathfrak{A}$-mesurables se conduisent comme des constantes vis-à-vis de l'opération $\mathsf{E}(\cdot \mid \mathfrak{A})$ (cf. propriété 1a).

6) *L'espérance mathématique conditionnelle vérifie toutes les inégalités fondamentales relatives à l'espérance mathématique ordinaire* et en particulier l'inégalité de Cauchy-Bouniakovski

$$\mathbf{E}(|\xi_1\xi_2| \mid \mathfrak{A}) \leqslant [\mathbf{E}(\xi_1^2 \mid \mathfrak{A})\mathbf{E}(\xi_2^2 \mid \mathfrak{A})]^{1/2}$$

et l'inégalité de Jensen : si $\mathbf{E}\,|\xi| < \infty$, alors pour toute fonction $g(x)$ convexe vers le bas on a

$$g(\mathbf{E}(\xi \mid \mathfrak{A})) \leqslant \mathbf{E}(g(\xi) \mid \mathfrak{A}).$$

7) *Formule des probabilités totales* (propriété 2 de la définition 2 pour $A = \Omega$)

$$\mathbf{E}\xi = \mathbf{E}\mathbf{E}(\xi \mid \mathfrak{A}).$$

8) *Moyennisation successive* (généralisation de la propriété 7)) : si $\mathfrak{A} \subset \subset \mathfrak{A}_1 \subset \mathfrak{F}$, alors

$$\mathbf{E}(\xi \mid \mathfrak{A}) = \mathbf{E}(\mathbf{E}(\xi \mid \mathfrak{A}_1) \mid \mathfrak{A}).$$

La démonstration de ces propriétés est accessible dans l'Annexe III.

Il est évident que les propriétés 1), 3), 4), 5), 7), 8) sont valables pour les variables aléatoires aussi bien scalaires que vectorielles. Signalons tout particulièrement la propriété suivante de l'espérance mathématique conditionnelle.

9) On sait que la fonction $\varphi(a) = \mathbf{E}(\xi - a)^2$ atteint son minimum pour $a = \mathbf{E}\xi$ (cf. par exemple [11]). Une propriété analogue est valable pour l'espérance mathématique conditionnelle : *la fonction* $\mathbf{E}(\xi - a(\omega))^2$ *atteint son minimum sur les fonctions* $a(\omega)$ $\mathfrak{A}$-*mesurables pour* $a(\omega) = \mathbf{E}(\xi \mid \mathfrak{A})$.

En effet, $\mathbf{E}(\xi - a(\omega))^2 = \mathbf{E}\mathbf{E}((\xi - a(\omega))^2 \mid \mathfrak{A})$. Mais $a(\omega)$ se conduit comme une constante vis-à-vis de l'opération $\mathbf{E}(\cdot \mid \mathfrak{A})$(cf. propriété 5)), donc

$$\mathbf{E}((\xi - a(\omega))^2 \mid \mathfrak{A}) = \mathbf{E}((\xi - \mathbf{E}(\xi \mid \mathfrak{A}))^2 \mid \mathfrak{A}) + \mathbf{E}((\mathbf{E}(\xi \mid \mathfrak{A}) - a(\omega))^2 \mid \mathfrak{A}),$$

et cette expression atteint son minimum pour $a(\omega) = \mathbf{E}(\xi \mid \mathfrak{A})$. Cette propriété peut tenir lieu de définition de l'espérance mathématique conditionnelle équivalente à la définition 2 et aux termes de laquelle $\mathbf{E}(\xi \mid \mathfrak{A})$ peut être traitée comme la « projection » de ξ sur $\mathfrak{A}$.

La propriété 9) admet la généralisation suivante au cas où $\xi = (\xi_1, \ldots, \xi_s)$ est un vecteur aléatoire de R^s.

9A) *Soient* $V = \|v_{ij}\|$ *une matrice semi-définie positive d'ordre s,* $a \in R^s$ *et*

$$\zeta(a) = (\xi - a)V(\xi - a)^T$$

(pour $V = E$ on a en particulier $\zeta(a) = |\xi - a|^2$). *Alors la fonction* $\mathbf{E}\zeta(a)$ *atteint son minimum sur la classe A des fonctions* $\mathfrak{A}$-*mesurables pour* $a(\omega) = \mathbf{E}(\xi \mid \mathfrak{A})$.

DÉMONSTRATION. Elle est calquée sur le cas scalaire. Posons $\alpha = E(\xi \mid \mathfrak{A})$. Alors $E\zeta(a) = EE(\zeta(a) \mid \mathfrak{A})$ et

$$E(\zeta(a) \mid \mathfrak{A}) = E((\xi - a)V(\xi - a)^T \mid \mathfrak{A}) =$$
$$= E((\xi - \alpha)V(\xi - \alpha)^T \mid \mathfrak{A}) + E((\alpha - a)V(\xi - \alpha)^T \mid \mathfrak{A}) +$$
$$+ E((\xi - \alpha)V(\alpha - a)^T \mid \mathfrak{A}) + E((\alpha - a)V(\alpha - a)^T \mid \mathfrak{A}). \quad (4)$$

Comme $\alpha - a$ est un vecteur $\mathfrak{A}$-mesurable, la propriété 5) nous donne

$$E((\alpha - a)V(\xi - \alpha)^T \mid \mathfrak{A}) = (\alpha - a)VE((\xi - \alpha)^T \mid \mathfrak{A}) = 0,$$
$$E((\xi - \alpha)V(\alpha - a)^T \mid \mathfrak{A}) = [E((\xi - \alpha) \mid \mathfrak{A})]V(\alpha - a)^T = 0.$$

Le dernier terme de (4) étant positif, et nul pour $a = \alpha$, on obtient ce qu'on voulait. ◄

§ 10. Distributions conditionnelles

Outre les espérances mathématiques conditionnelles, on peut envisager les distributions conditionnelles par rapport à des sous-tribus et à des variables aléatoires. Dans ce paragraphe on étudiera les distributions conditionnelles par rapport à des variables aléatoires.

Soient ξ et η des variables aléatoires sur $(\Omega, \mathfrak{F}, P)$ à valeurs respectivement dans R^s et R^k, et soit $\mathfrak{B}^s$ la tribu des boréliens de R^s.

DÉFINITION 1. On dit qu'une fonction $P(B \mid y)$ des variables $y \in R^k$ et $B \in \mathfrak{B}^s$ est une *distribution conditionnelle de ξ par rapport à la condition $\eta = y$* si

1) pour tout B la fonction $P(B \mid \eta)$ est la probabilité conditionnelle $P(\xi \in B \mid \eta)$ de l'événement $\{\xi \in B\}$ par rapport à η, autrement dit $P(B \mid y)$ est une fonction borélienne de y telle que pour tout $A \in \mathfrak{B}^k$

$$E(P(B \mid \eta) ; \eta \in A) = \int_A P(B \mid y)P(\eta \in dy) = P(\xi \in B, \eta \in A) ;$$

2) $P(B \mid y)$ est la distribution des probabilités sur B pour tout y.

Nous écrirons parfois la fonction $P(B \mid y)$ sous la forme plus détaillée

$$P(B \mid y) = P(\xi \in B \mid \eta = y).$$

Nous savons que pour tout $B \in \mathfrak{B}^s$ il existe une fonction borélienne $g_B(y)$ telle que $g_B(\eta) = P(\xi \in B \mid \eta)$. En posant $P(B \mid y) = g_B(y)$ on satisfait la condition 1) de la définition 1. Mais alors la condition 2) ne découle en aucune façon des propriétés de l'espérance mathématique conditionnelle et ne doit pas être nécessairement réalisée : en effet, la probabilité conditionnelle $P(\xi \in B \mid \eta)$ est définie pour tout B aux valeurs près sur un ensemble de mesure nulle N_B (de sorte qu'il existe plusieurs variantes d'espérances

mathématiques conditionnelles) et cet ensemble peut varier d'un ensemble B à un autre. Donc, si la réunion $N = \bigcup_{B\in\mathfrak{B}^s} N_B$ a une probabilité non nulle, il est possible que, par exemple, pour aucun ω de N l'égalité

$$\mathbf{P}(\xi \in B_1 \cup B_2 | \eta) = \mathbf{P}(\xi \in B_1 | \eta) + \mathbf{P}(\xi \in B_2 | \eta)$$

(qui exprime l'additivité de la probabilité) ne soit réalisée simultanément pour tous les B_1 et B_2 disjoints de $\mathfrak{B}^s$, c'est-à-dire que la fonction $g_B(y)$ ne sera pas une distribution (en tant que fonction de B) sur un ω-ensemble N de probabilité strictement positive.

Mais *dans notre cas où ξ est une variable aléatoire à valeurs dans R^s muni de la tribu des boréliens $\mathfrak{B}^s$, la fonction $g_B(\eta) = \mathbf{P}(\xi \in B | \eta)$ peut toujours être choisie de telle sorte que $g_B(y)$ soit une distribution conditionnelle* (cf. [24], [34]).

Comme il fallait s'y attendre les espérances mathématiques conditionnelles s'expriment par des intégrales par rapport aux distributions conditionnelles.

THÉORÈME 1. *Pour toute fonction mesurable $g(x)$ de R^s dans R telle que $\mathbf{E}|g(\xi)| < \infty$, on a l'égalité*

$$\mathbf{E}(g(\xi)|\eta) = \int g(x) P(dx|\eta). \tag{1}$$

DÉMONSTRATION. Il suffit de traiter le cas où $g(x) \geqslant 0$. Si $g(x) = I_A(x)$ est l'indicateur de l'ensemble A, la formule (1) est manifestement vérifiée. Donc, elle est valable pour toute fonction simple $g_n(x)$ (c'est-à-dire pour toute fonction prenant un nombre fini de valeurs). Reste à considérer une suite $g_n \uparrow g$ et à se servir de la monotonie des deux membres de (1) et de la propriété 4) du § 9. ◄

Pour calculer les distributions conditionnelles, on peut se servir de la règle élémentaire suivante que, pour plus de suggestion, nous écrirons sous la forme

$$\mathbf{P}(\xi \in B | \eta = y) = \frac{\mathbf{P}(\xi \in B, \eta \in dy)}{\mathbf{P}(\eta \in dy)}. \tag{2}$$

Il est évident que les deux conditions de la définition 1 sont formellement remplies.

Si ξ et η admettent des densités de probabilité, cette égalité acquiert une signification exacte.

DÉFINITION 2. Supposons que la distribution conditionnelle $P(B|y)$ est, pour tout y, absolument continue par rapport à une mesure μ dans R^s :

$$\mathbf{P}(\xi \in B | \eta = y) = \int_B f(x|y) \mu(dx).$$

Alors la densité $f(x|y)$ s'appelle *densité conditionnelle de* ξ *(par rapport à la mesure* μ*) sachant que* $\eta = y$.

En d'autres termes, la fonction $f(x|y)$ mesurable par rapport à x et y est densité conditionnelle de ξ sachant que $\eta = y$ si

1) pour tous boréliens $A \subset R^k$, $B \subset R^s$

$$\int_{y\in A}\int_{x\in B} f(x|y)\mu(dx)\mathbf{P}(\eta \in dy) = \mathbf{P}(\xi \in B, \eta \in A)\,; \qquad (3)$$

2) la fonction $f(x|y)$ est une densité de probabilité pour tout y.

Du théorème 1 il s'ensuit que si existe la densité conditionnelle, on a

$$\mathbf{E}(g(\xi)|\eta) = \int g(x)f(x|\eta)\mu(dx).$$

Si l'on admet accessoirement que la distribution de η admet la densité $q(y)$ par rapport à une mesure λ dans R^k, alors la relation (3) peut être mise sous la forme

$$\int_{y\in A}\int_{x\in B} f(x|y)q(y)\mu(dx)\lambda(dy) = \mathbf{P}(\xi \in B, \eta \in A). \qquad (4)$$

Considérons maintenant le produit direct des espaces R^s et R^k, et sur ce produit le produit direct des mesures $\mu \times \lambda$ (si $C = B \times A$, $B \subset R^s$, $A \subset R^k$, alors $\mu \times \lambda(C) = \mu(B)\lambda(A)$). La relation (4) exprime de toute évidence que la distribution conjointe de ξ et de η dans $R^s \times R^k$ admet par rapport à $\mu \times \lambda$ une densité égale à

$$f(x, y) = f(x|y)q(y).$$

On a le théorème réciproque.

THÉORÈME 2. *Si la distribution conjointe de* ξ *et de* η *dans* $R^s \times R^k$ *admet une densité* $f(x, y)$ *par rapport à* $\mu \times \lambda$, *alors la fonction*

$$f(x|y) = \frac{f(x, y)}{q(y)}, \quad \text{où} \quad q(y) = \int f(x, y)\mu(dx),$$

est la densité conditionnelle de ξ *sachant que* $\eta = y$, *et la fonction* $q(y)$, *la densité de* η *par rapport à la mesure* λ.

DÉMONSTRATION. Relativement à $q(y)$ cette proposition est évidente, puisque $\int_A q(y)\lambda(dy) = \mathbf{P}(\eta \in A)$. Reste à remarquer que $f(x|y) = f(x, y)/q(y)$ satisfait toutes les conditions de la définition 2 de la densité conditionnelle (l'égalité (4) qui est identique à (3) est réalisée de façon évidente). ◄

REMARQUE 1. Les variables aléatoires ξ et η peuvent être permutées dans le théorème 2. Dans ces conditions, outre $f(x|y)$ il existe la densité

conditionnelle

$$q(y|x) = \frac{f(x, y)}{f(x)}, \quad f(x) = \int f(x, y)\lambda(dy),$$

de la variable aléatoire η sachant que $\xi = x$. Cette conséquence élémentaire du théorème 2 jouera un rôle important dans la suite de l'exposé. Appliquée aux problèmes de statistique, elle nous permettra d'établir la formule de Bayes qui sera utilisée tout au long de cet ouvrage.

EXEMPLE 1. Soit Φ_{α, σ^2} la distribution normale à deux dimensions des variables ξ_1 et ξ_2, où $\alpha = (\alpha_1, \alpha_2)$, $\alpha_i = \mathbf{E}\xi_i$, $\sigma^2 = \|\sigma_{ij}\|$, $\sigma_{ij} = \mathbf{E}(\xi_i - \alpha_i)(\xi_j - \alpha_j)$, $i, j = 1, 2$. Le déterminant de la matrice des moments d'ordre deux est égal à

$$|\sigma^2| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \varrho^2),$$

où ϱ est le coefficient de corrélation entre ξ_1 et ξ_2. Donc, si $|\varrho| \neq 1$, la matrice des moments d'ordre deux n'est pas dégénérée et son inverse est

$$A = (\sigma^2)^{-1} = \frac{1}{|\sigma^2|}\begin{Vmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{Vmatrix} =$$

$$= \frac{1}{1-\varrho^2}\begin{Vmatrix} \dfrac{1}{\sigma_{11}} & -\dfrac{\varrho}{\sqrt{\sigma_{11}\sigma_{22}}} \\ -\dfrac{\varrho}{\sqrt{\sigma_{11}\sigma_{22}}} & \dfrac{1}{\sigma_{22}} \end{Vmatrix}.$$

Par conséquent, la densité conjointe de ξ_1 et de ξ_2 (par rapport à la mesure de Lebesgue) est égale à (cf. § 2)

$$f(x, y) = \frac{1}{2\pi\sigma_{11}\sigma_{22}\sqrt{1-\varrho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\varrho^2)}\left[\frac{(x-\alpha_1)^2}{\sigma_{11}} - \frac{2\varrho(x-\alpha_1)(y-\alpha_2)}{\sqrt{\sigma_{11}\sigma_{22}}} + \frac{(y-\alpha_2)^2}{\sigma_{22}}\right]\right\}.$$

Les densités de ξ_1 et de ξ_2 sont respectivement égales à

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_{11}}} e^{-\frac{(x-\alpha_1)^2}{2\sigma_{11}}},$$

$$q(y) = \frac{1}{\sqrt{2\pi\sigma_{22}}} e^{-\frac{(y-\alpha_2)^2}{2\sigma_{22}}}.$$

Donc, la densité conditionnelle de ξ_1 sachant que $\xi_2 = y$ est

$$f(x|y) = \frac{f(x, y)}{d(y)} =$$

$$= \frac{1}{\sqrt{2\pi\sigma_{11}(1-\varrho^2)}} \exp\left\{-\frac{1}{2\sigma_{11}(1-\varrho^2)}\left(x-\alpha_1-\varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(y-\alpha_2)\right)^2\right\};$$

on reconnaît ici la densité de la distribution normale de moyenne $\alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}(y - \alpha_2)$ et de variance $\sigma_{11}(1 - \varrho^2)$. De là il s'ensuit en particulier que l'espérance mathématique conditionnelle de ξ_1 par rapport à ξ_2 est

$$\mathbf{E}(\xi_1|\xi_2) = \alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}\,(\xi_2 - \alpha_2).$$

La droite $x = \alpha_1 + \varrho\sqrt{\frac{\sigma_{11}}{\sigma_{22}}}(y - \alpha_2)$ s'appelle *droite de régression* de ξ_1 en ξ_2. Elle donne les meilleures approximations en moyenne quadratique de ξ_1 sachant que $\xi_2 = y$.

EXEMPLE 2. Soit à calculer la densité d'une variable aléatoire $\xi = \varphi(\zeta, \eta)$, où ζ et η sont indépendantes. De la formule (3) pour $A = R^k$ il résulte que la densité $f(x)$ de la distribution de ξ s'exprime en fonction de la densité conditionnelle $f(x|y)$ par l'égalité

$$f(x) = \int f(x|y)\mathbf{P}(\eta \in dy). \tag{5}$$

Dans le problème posé, par $f(x|y)$ il faut comprendre la densité de la variable aléatoire $\varphi(\zeta, y)$, puisque $\mathbf{P}(\xi \in B|\eta = y) = \mathbf{P}(\varphi(\zeta, y) \in B)$.

La formule (5) est très utile pour le calcul des distributions des statistiques. Au n°6 du § 2 par exemple, on aurait pu écrire directement la formule (2.7) de la densité de la distribution de Fisher sans la déduire de la forme de la fonction de répartition.

§ 11. Approches bayésienne et minimax de l'estimation des paramètres

Le principe de l'approche bayésienne consiste à traiter le paramètre inconnu θ comme une *variable aléatoire* admettant une densité de probabilité $q(t)$, $t \in \Theta$, (connue ou non) par rapport à une mesure λ qui, comme la mesure μ de la condition (A_μ), sera soit la mesure de Lebesgue, soit une mesure discrète. La densité $q(t)$ s'appelle *densité a priori*. L'approche bayésienne suppose que le paramètre inconnu θ est un paramètre aléatoire de densité de probabilité $q(t)$.

Supposons par ailleurs que $f_t(x)$, $t \in \Theta$, $x \in \mathscr{X}^n$, est la fonction de vraisemblance introduite au § 6. Comme déjà signalé, pour chaque t la fonction

$f_t(x)$ est une densité de probabilité dans $\mathcal{X}^n$. Donc, la fonction

$$f(x, t) = f_t(x)q(t)$$

est la densité d'une distribution dans $\mathcal{X}^n \times \Theta$ par rapport à la mesure $\mu^n \times \lambda$, qui peut être interprétée comme la *densité de la distribution conjointe* de X et de θ. Dans cette approche, le théorème 10.2 nous dit que la fonction $f_t(x)$, $x \in \mathcal{X}^n$, est la *densité conditionnelle de X sachant que $\theta = t$* :

$$f_t(x) = f(x|t), \quad \mathsf{E}_\theta g(X) = \mathsf{E}(g(X)|\theta).$$

Dans ces considérations, l'aspect formel des choses implique que $f_t(x)$ soit mesurable par rapport à t et à x. On admettra qu'il en est ainsi partout où cela est nécessaire.

Dans la suite, le paramètre sera désigné par θ s'il est traité comme une variable aléatoire, et par t, u, etc., s'il est fixé, de sorte que

$$\mathsf{E}_t g(X) = \mathsf{E}(g(X)|\theta = t).$$

On peut, en plus de $f(x|t)$, écrire la formule de la *densité conditionnelle $q(t|x)$ de la variable θ sachant que $X = x$* :

$$q(t|x) = \frac{f_t(x)q(t)}{f(x)}, \quad f(x) = \int f_t(x)q(t)\lambda(dt). \tag{1}$$

Cette densité définit la distribution *a posteriori* de θ que l'on désignera par $\mathbf{Q}_x$. L'égalité (1) s'appelle *formule de Bayes* de la densité de la distribution *a posteriori*. Cette formule jouera un rôle important dans la suite de l'exposé.

Dans l'approche bayésienne, la propriété 9 de l'espérance mathématique conditionnelle exprime que parmi les fonctions $\theta^* = \varphi(X)$ le meilleur estimateur de θ (au sens de la minimisation de $\mathsf{E}(\theta - \varphi(X))^2$) est la fonction

$$\theta^*_{\mathbf{Q}} = \mathsf{E}(\theta|X) = \int tq(t|X)\lambda(dt) = \int t\mathbf{Q}_X(dt). \tag{2}$$

DÉFINITION 1. L'estimateur $\theta^*_{\mathbf{Q}}$ défini par les formules (2), (1) s'appelle *estimateur bayésien associé à la distribution a priori* $\mathbf{Q}$ *de densité* $q(t)$.

Signalons encore que la dispersion quadratique moyenne

$$\mathsf{E}(\theta^* - \theta)^2 = \mathsf{E}\mathsf{E}((\theta^* - \theta)^2|\theta) = \mathsf{E}\mathsf{E}_\theta(\theta^* - \theta)^2 = \\ = \int \mathsf{E}_t(\theta^* - t)^2 q(t)\lambda(dt) \tag{3}$$

atteint son minimum pour un estimateur bayésien. La relation (3) montre qu'un estimateur bayésien minimise la valeur moyenne (de fonction de poids $q(t)\lambda(dt)$ donnée) de $\mathsf{E}_t(\theta^* - t)^2$.

Autrement dit, si θ est un paramètre aléatoire de densité $q(t)$, *le meilleur estimateur au sens de l'approche de la moyenne quadratique est un estima-*

teur bayésien. La dispersion quadratique moyenne (3) d'un estimateur bayésien peut être mise sous la forme (cf. (1))

$$\mathbf{E}(\theta^*_{\mathrm{Q}} - \theta)^2 = \int \mathbf{E}_t(\theta^*_{\mathrm{Q}} - t)^2 q(t)\lambda(dt) =$$
$$= \int\int (t - \theta^*_{\mathrm{Q}})^2 f_t(x) q(t)\lambda(dt)\mu^n(dx) = \int \sigma^2_{\mathrm{Q}x} f(x)\mu^n(dx) = \mathbf{E}\sigma^2_{\mathrm{Q}X},$$

où $\sigma^2_{\mathrm{Q}X}$ est la variance de la distribution *a posteriori* $\mathbf{Q}_X$:

$$\sigma^2_{\mathrm{Q}X} = \int (t - \theta^*_{\mathrm{Q}})^2 q(t\,|\,X)\lambda(dt) = \int (t - \mathbf{E}(\theta\,|\,X))^2 \mathbf{Q}_X(dt). \tag{4}$$

L'autre méthode de comparaison des estimateurs, mentionnée au § 8, consiste à comparer $\sup_{t\in\Gamma} \mathbf{E}_t(\theta^* - t)^2$, où $\Gamma \in \Theta$ est un sous-ensemble donné de Θ (Γ est confondu soit avec Θ, soit avec la partie de Θ au sujet de laquelle on a réussi à établir que $\theta \in \Gamma$).

DÉFINITION 2. Un estimateur $\bar{\theta}^*$ est dit *minimax* si pour tout autre estimateur θ^* on a

$$\sup_{t\in\Gamma} \mathbf{E}_t(\bar{\theta}^* - t)^2 \leqslant \sup_{t\in\Gamma} \mathbf{E}_t(\theta^* - t)^2.$$

En d'autres termes, pour un estimateur minimax on a

$$\inf_{\theta^*} \sup_{t\in\Gamma} \mathbf{E}_t(\theta^* - t)^2 = \sup_{t\in\Gamma} \mathbf{E}_t(\bar{\theta}^* - t)^2. \tag{5}$$

Etablissons quelques relations utiles entre estimateurs bayésiens et minimax.

THÉORÈME 1. *Désignons par θ^*_{Q} l'estimateur bayésien associé à une distribution a priori* $\mathbf{Q}$ *de densité q. S'il existe un estimateur θ^*_1 et une distribution* $\mathbf{Q}$ *tels que pour tous les t*

$$\mathbf{E}_t(\theta^*_1 - t)^2 \leqslant \int \mathbf{E}_u(\theta^*_{\mathrm{Q}} - u)^2 q(u)\lambda(du), \tag{6}$$

alors l'estimateur θ^*_1 est minimax.

DÉMONSTRATION. Soit θ^* un autre estimateur. Alors

$$\sup_t \mathbf{E}_t(\theta^* - t)^2 \geqslant \int \mathbf{E}_t(\theta^* - t)^2 q(t)\lambda(dt) \geqslant$$
$$\geqslant \int \mathbf{E}_t(\theta^*_{\mathrm{Q}} - t)^2 q(t)\lambda(dt) \geqslant \mathbf{E}_t(\theta^*_1 - t)^2. \blacktriangleleft$$

A noter que dans la relation (6) l'égalité est nécessairement réalisée pour presque tous les t appartenant au support $N_{\mathrm{Q}} = \{t : q(t) > 0\}$ de la distribution $\mathbf{Q}$, puisque, le cas échéant, on aurait

$$\int \mathbf{E}_t(\theta^*_1 - t)^2 q(t)\lambda(dt) < \int \mathbf{E}_t(\theta^*_{\mathrm{Q}} - t)^2 q(t)\lambda(dt),$$

ce qui contredit la définition d'un estimateur bayésien.

Cette remarque nous permet de formuler le critère minimax suivant d'un estimateur, équivalent au théorème 1.

THÉORÈME 2. *Un estimateur θ^* est minimax si*
1) *il est bayésien pour une distribution* $\mathbf{Q}$,
2) $\mathbf{E}_t(\theta^* - t)^2 = c = \text{const}$ *pour* $t \in N_{\mathbf{Q}}$,
3) $\mathbf{E}_t(\theta^* - t)^2 \leqslant c$ *pour les autres* t.

Si $\theta^* = \theta^*_{\mathbf{Q}} = \bar{\theta}^*$ vérifie ce critère, il est évident que

$$\sup_t \mathbf{E}_t(\bar{\theta}^* - t)^2 = \int \mathbf{E}_t(\bar{\theta}^* - t)^2 q(t)\lambda(dt). \tag{7}$$

Donc, un estimateur minimax est un estimateur bayésien qui « lisse » les erreurs $\mathbf{E}_t(\bar{\theta}^* - t)^2$ pour divers t. Cela signifie que la distribution *a priori* $\overline{\mathbf{Q}}$ associée à cet estimateur accorde la même importance à toutes les valeurs possibles de θ et ne privilégie pas certaines valeurs (les plus probables) de θ comme le font les estimateurs bayésiens $\theta^*_{\mathbf{Q}}$ associés à d'autres distributions *a priori* $\mathbf{Q} \neq \overline{\mathbf{Q}}$. Vu que dans le dernier cas on s'est servi d'une information supplémentaire sur θ, il est naturel que pour $\mathbf{Q} \neq \overline{\mathbf{Q}}$ les estimateurs $\theta^*_{\mathbf{Q}}$ soient tels que

$$\int \mathbf{E}_t(\theta^*_{\mathbf{Q}} - t)^2 \mathbf{Q}(dt) \leqslant \int \mathbf{E}_t(\theta^*_{\overline{\mathbf{Q}}} - t)^2 \overline{\mathbf{Q}}(dt).$$

Pour cette raison, la distribution $\overline{\mathbf{Q}}$ associée à l'estimateur minimax $\bar{\theta}^*$ est souvent dite *la plus défavorable.*

Etant donné que la distribution $\overline{\mathbf{Q}}$ n'existe pas toujours (c'est généralement le cas lorsque Θ n'est pas borné), on peut se servir du critère modifié suivant pour déterminer un estimateur minimax.

THÉORÈME 3. *S'il existe un estimateur θ^*_1 et une suite de distributions $\mathbf{Q}^{(k)}$ de densités $q^{(k)}$ tels que pour tous les t*

$$\mathbf{E}_t(\theta^*_1 - t)^2 \leqslant \operatorname*{lim.sup}_{k\to\infty} \int \mathbf{E}_t(\theta^*_{\mathbf{Q}^{(k)}} - t)^2 q^{(k)}(t)\lambda(dt),$$

alors θ^*_1 est un estimateur minimax.

DÉMONSTRATION. Elle coule de source. Pour tout estimateur θ^*, on a

$$\sup_t \mathbf{E}_t(\theta^* - t)^2 \geqslant \int \mathbf{E}_t(\theta^* - t)^2 q^{(k)}(t)\lambda(dt) \geqslant \int \mathbf{E}_t(\theta^*_{\mathbf{Q}^{(k)}} - t)^2 q^{(k)}(t)\lambda(dt).$$

D'où

$$\sup_t \mathbf{E}_t(\theta^* - t)^2 \geqslant \operatorname*{lim.sup}_{k\to\infty} \int \mathbf{E}_t(\theta^*_{\mathbf{Q}^{(k)}} - t)^2 q^{(k)}(t)\lambda(dt) \geqslant \mathbf{E}_t(\theta^*_1 - t)^2. \blacktriangleleft$$

EXEMPLE 1. Soit $X \in \Phi_{\alpha,1}$, et soit à déterminer un estimateur bayésien $\alpha^*_{Q^{(k)}}$ du paramètre α qui suit une distribution normale *a priori* $Q^{(k)} = \Phi_{(0,k)}$. Dans ce cas on doit poser $\lambda(dt) = dt$,

$$q^{(k)}(t) = \frac{1}{\sqrt{2\pi k}} e^{-\frac{t^2}{2k}}.$$

La distribution *a posteriori* $Q_X^{(k)}$ admettra une densité $q^{(k)}(t \mid X)$ proportionnelle (comme fonction de t) à $q^{(k)}(t) f_t(X)$ ou, ce qui revient au même, proportionnelle à

$$\exp\left\{-\frac{t^2}{2k} - \frac{1}{2}\sum (x_i - t)^2\right\}.$$

L'égalité

$$-\frac{t^2}{2}\left(\frac{1}{k} + n\right) + \bar{x}nt = -\frac{1}{2}\left(\frac{1}{k} + n\right)\left(t - \frac{\bar{x}n}{\frac{1}{k} + n}\right)^2 + \frac{(\bar{x}n)^2}{2\left(\frac{1}{k} + n\right)}$$

entraîne

$$Q_X^{(k)} = \Phi_{\frac{\bar{x}nk}{1+nk},\ \frac{k}{1+nk}}.$$

Comme l'estimateur bayésien $\alpha^*_{Q^{(k)}}$ du paramètre α est égal à l'espérance mathématique de la distribution *a posteriori*, on en déduit que

$$\alpha^*_{Q^{(k)}} = \frac{\bar{x}nk}{1 + nk} = \frac{\bar{x}}{1 + \frac{1}{nk}}.$$

La variance $\sigma^2_{Q_X^{(k)}} = \dfrac{k}{1 + nk}$ de la distribution *a posteriori* ne dépend pas de X. Donc, en vertu de (4), l'erreur quadratique moyenne de l'estimateur bayésien est égale à $\dfrac{k}{1 + nk}$ et tend vers $\dfrac{1}{n}$ lorsque $k \to \infty$. Donc, pour $\alpha^* = \bar{x}$, on a

$$E_t(\bar{x} - t)^2 = \frac{1}{n} = \lim_{k \to \infty} \int E_t(\alpha^*_{Q^{(k)}} - t)^2 q^{(k)}(t)\,dt,$$

et par suite, cet estimateur est minimax en vertu du théorème 3. La distribution « la plus défavorable » aurait été la distribution uniforme sur la droite tout entière (la distribution « limite » de $\Phi_{0,k}$) si elle eût existé *).

*) Il est intéressant de signaler que l'estimateur $\alpha^* = \bar{x}$ ne jouit plus de cette propriété si X est un échantillon issu d'une loi normale multidimensionnelle de dimension > 2 ($x_j \in R^k$, $\alpha \in R^k$, $k \geqslant 3$). Pour plus de détails voir [42].

Dans l'exemple suivant, l'ensemble Θ est compact et la distribution « la plus défavorable » existe.

EXEMPLE 2. Soit $X \in \mathbf{B}_p$, c'est-à-dire que x_j, $j = 1, \ldots, n$, prennent les valeurs 1 et 0 avec les probabilités respectives p et $1-p$, $p \in \Theta = [0, 1]$. On sait que dans ce cas, pour l'estimateur $p^* = \bar{x}$ on a

$$\mathsf{E}_p(\bar{x} - p)^2 = p(1 - p)/n,$$

de sorte que le critère du théorème 2 n'est pas rempli. Considérons l'estimateur

$$p^* = \frac{\bar{x} + \dfrac{1}{2\sqrt{n}}}{1 + \dfrac{1}{\sqrt{n}}}. \tag{8}$$

Son erreur quadratique moyenne

$$\mathsf{E}_p(p^* - p)^2 = \left(1 + \frac{1}{\sqrt{n}}\right)^{-2} \mathsf{E}_p\left(\bar{x} - p + \frac{1}{2\sqrt{n}} - \frac{p}{\sqrt{n}}\right)^2 =$$

$$= \frac{n}{(1 + \sqrt{n})^2}\left(\frac{p(1 - p)}{n} + \frac{(1 - 2p)^2}{4n}\right) = \frac{1}{4(1 + \sqrt{n})^2}$$

ne dépend pas de p. Si l'on s'assure maintenant que l'estimateur (8) est bayésien, on démontre *ipso facto* qu'il est minimax. Considérons la distribution *a priori* $\mathbf{Q} = \mathbf{B}_{N+1,\, N+1}$, où $\mathbf{B}_{\lambda_1,\, \lambda_2}$ est une distribution bêta de densité (cf. n°8 du § 2)

$$\frac{\Gamma(\lambda_1 + \lambda_2)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\, t^{\lambda_1 - 1}(1 - t)^{\lambda_2 - 1}.$$

Puisque

$$f_t(X) = t^{\bar{x}n}(1 - t)^{n(1 - \bar{x})},$$

$$q(t) = \frac{\Gamma(2N + 2)}{\Gamma^2(N + 1)}\, t^N(1 - t)^N,$$

la distribution *a posteriori* admettra la densité $q(t \mid X)$ qui, en tant que fonction de t, sera proportionnelle à $f_t(X)q(t)$ ou, ce qui revient au même, à $t^{N + \bar{x}n}(1 - t)^{N + (1 - \bar{x})n}$. Ceci exprime que la distribution *a posteriori* est confondue avec $\mathbf{B}_{N + \bar{x}n + 1,\, N + n(1 - \bar{x}) + 1}$. La moyenne de la distribution $\mathbf{B}_{\lambda_1,\, \lambda_2}$ étant égale à $\lambda_1/(\lambda_1 + \lambda_2)$ (cf. n°8 du § 2), l'estimateur bayésien $p^*_{\mathbf{Q}}$ associé à $\mathbf{Q}$ sera

$$p^*_{\mathbf{Q}} = \frac{N + \bar{x}n + 1}{2N + n + 2} = \frac{\bar{x} + (N + 1)/n}{1 + 2(N + 1)/n}.$$

Pour $N + 1 = \sqrt{n}/2$, cet estimateur s'identifie à l'estimateur p^* défini dans (8) et sera minimax en vertu du théorème 2. La distribution $\mathbf{Q}$ sera la plus défavorable. Lorsque n croît, elle se concentre autour de la « plus mauvaise » des valeurs du paramètre p, la valeur 1/2 pour laquelle la variance de l'estimateur $\bar{x}$, qui est égale à $p(1 - p)/n = 1/(4n)$, sera maximale. L'estimateur $\bar{x}$ n'est pas minimax, puisque

$$\sup_p \frac{p(1 - p)}{n} = \frac{1}{4n} > \frac{1}{4(1 + \sqrt{n})^2} .$$

Dans le même temps il est clair que pour toutes les valeurs de p, extérieures à un voisinage étroit de $p = 1/2$, l'estimateur $\bar{x}$ sera meilleur que $p_{\mathbf{Q}}^*$: ceci aura lieu pour tous les p tels que

$$p(1 - p) < \frac{1}{4(1 + 1/\sqrt{n})^2} .$$

Dans le cas général, il n'est pas toujours possible d'expliciter (par des fonctions de X) les estimateurs bayésiens et minimax. L'approche asymptotique s'impose alors tout naturellement.

Avant d'introduire les définitions respectives, rappelons que les estimateurs bayésien et minimax $\theta_{\mathbf{Q}}^*$ et $\bar{\theta}^*$ ont été définis par les inégalités

$$\begin{gathered} \mathbf{E}(\theta_{\mathbf{Q}}^* - \theta)^2 - \mathbf{E}(\theta^* - \theta)^2 \leqslant 0, \\ \sup_{t \in \Gamma} \mathbf{E}_t(\bar{\theta}^* - t)^2 - \sup_{t \in \Gamma} \mathbf{E}_t(\theta^* - t)^2 \leqslant 0 \end{gathered} \tag{9}$$

pour tout estimateur θ^*. Il aurait été contraire à la raison de définir les estimateurs asymptotiquement bayésien et asymptotiquement minimax en ajoutant simplement aux premiers membres le signe du passage à la limite $\lim_{n\to\infty}$, puisque généralement pour les estimateurs asymptotiquement normaux, $\mathbf{E}_\theta(\theta^* - \theta)^2 \sim \sigma^2(\theta)/n$ et les premiers membres de (9) tendront vers 0. Il est donc naturel d'envisager, disons, le rapport des termes de (9). Vu que dans la suite nous aurons essentiellement affaire à des estimateurs pour lesquels $\mathbf{E}_\theta(\theta^* - \theta)^2$ sera de l'ordre de $1/n$, il est équivalent de se servir de la définition suivante.

DÉFINITION 3. On dit qu'un estimateur θ_1^* est *asymptotiquement bayésien* ou *asymptotiquement minimax* si pour tout autre estimateur θ^*, on a respectivement

$$\limsup_{n\to\infty} [\mathbf{E}_n(\theta_1^* - \theta)^2 - \mathbf{E}_n(\theta^* - \theta)^2] \leqslant 0,$$

$$\limsup_{n\to\infty} [\sup_{t\in\Gamma} \mathbf{E}_t n(\theta_1^* - t)^2 - \sup_{t\in\Gamma} \mathbf{E}_t n(\theta^* - t)^2] \leqslant 0.$$

Nous verrons qu'il est toujours possible de trouver des estimateurs asymptotiquement bayésiens et asymptotiquement minimax sous des conditions assez peu contraignantes.

Dans le *cas multidimensionnel* (c'est-à-dire lorsque $\theta \in R^k$ est un vecteur) la propriété 9) de l'espérance mathématique conditionnelle reste en vigueur et l'estimateur

$$\theta_Q^* = \mathsf{E}(\theta \mid X)$$

minimisera

$$v(\theta^*) = \mathsf{E}(\theta^* - \theta)V(\theta^* - \theta)^T = \mathsf{E}\mathsf{E}_\theta(\theta^* - \theta)V(\theta^* - \theta)^T =$$
$$= \int \mathsf{E}_t(\theta^* - t)V(\theta^* - t)^T q(t)\lambda(dt)$$

pour toute matrice semi-définie positive V ou ce qui est équivalent (cf. § 8) minimisera la moyenne (avec le poids $q(t)$) de la dispersion quadratique moyenne de $\theta^* - \theta$ suivant toute direction $a \in R^k$.

DÉFINITION 4. On dit qu'un estimateur θ_Q^* est *bayésien* si pour tout autre estimateur θ^* et toute matrice semi-définie positive V on a

$$v(\theta_Q^*) \leqslant v(\theta^*)$$

Un estimateur θ_1^* est *asymptotiquement bayésien* si

$$\limsup_{n\to\infty} [nv(\theta_1^*) - nv(\theta_Q^*)] \leqslant 0.$$

DÉFINITION 5. Un estimateur $\bar\theta^*$ est *minimax* si pour tout autre estimateur θ^* et toute matrice semi-définie positive V on a

$$\sup_{t\in\Gamma} \mathsf{E}_t(\bar\theta^* - t)V(\bar\theta^* - t)^T - \sup_{t\in\Gamma} \mathsf{E}_t(\theta^* - t)V(\theta^* - t)^T \leqslant 0.$$

Un estimateur θ_1^* est *asymptotiquement minimax* si

$$\limsup_{n\to\infty} [\sup_{t\in\Gamma} \mathsf{E}_t n(\theta_1^* - t)V(\theta_1^* - t)^T - \sup_{t\in\Gamma} \mathsf{E}_t n(\bar\theta^* - t)V(\bar\theta^* - t)^T] \leqslant 0.$$

Signalons encore une fois en conclusion de ce paragraphe que dans le cas bayésien on peut au besoin traiter $\mathsf{E}_\theta S$, $\mathbf{P}_\theta(A)$ et $f_\theta(x)$ d'un point de vue nouveau, plus exactement comme l'espérance mathématique $\mathsf{E}(S \mid \theta)$, la probabilité $\mathbf{P}(A \mid \theta)$ et la densité $f(x \mid \theta)$, conditionnelles par rapport à θ.

§ 12. Statistiques exhaustives

Dans le paragraphe précédent nous avons examiné la construction de deux types d'estimateurs optimaux : les estimateurs bayésiens et les estimateurs minimax. On se propose d'introduire ici la notion de statistique exhaustive qui nous permettra de construire des estimateurs efficaces (cf. § 8).

Les statistiques exhaustives jouent un rôle important en statistique mathématique en général et en théorie de l'estimation en particulier.

Convenons de désigner par $S = S(X)$ les statistiques qui sont des fonctions mesurables (scalaires ou vectorielles) de X.

Soient $X \in \mathbf{P}_\theta$, $\mathbf{P}_\theta \in \mathscr{P} = \{\mathbf{P}_\theta\}$. Considérons la distribution $\mathbf{P}_\theta(X \in B|S)$, $B \in \mathfrak{B}^n_{\mathscr{X}}$; conditionnelle par rapport à la variable aléatoire S, engendrée par une distribution $\mathbf{P}_\theta$ dans $\mathscr{X}^n$.

DÉFINITION 1. On dit qu'une statistique $S = S(X)$ est *exhaustive pour le paramètre* θ s'il existe une distribution conditionnelle $\mathbf{P}_\theta(X \in B|S)$ indépendante de θ.

On sait que $\mathbf{P}_\theta(X \in B|S)$ est une espérance mathématique conditionnelle pour tout B. Il existe donc une fonction $P(B|s)$, borélienne par rapport à s pour tout B, telle que

$$\mathbf{P}_\theta(X \in B|S) = P(B|S).$$

On peut admettre (cf. § 10) que $P(B|s)$, traitée comme une fonction de B, est une *distribution conditionnelle sachant que* $S = s$. Cette distribution peut être interprétée comme une *distribution de X sur la surface* $S(\mathbf{x}) = s$.

Mais si S est une statistique exhaustive, c'est que cette distribution est *indépendante* de θ! Cela signifie que la connaissance de la position du point échantillon X sur la surface $S(\mathbf{x}) = s$ ne nous fournit aucune information supplémentaire sur le paramètre θ. (En effet, il est clair que personne ne s'aventurera à déterminer le paramètre inconnu θ dans l'exemple 1 de l'Introduction en jetant une pièce de monnaie, pour la raison simple que la distribution du nombre de « piles » ou de « faces » ne dépend en aucune façon de θ.)

Cette circonstance signifie à son tour que toute l'information sur le paramètre θ est contenue dans la valeur de la statistique S. D'où son nom de statistique exhaustive : *grosso modo*, la connaissance de $S(X)$ *suffit* pour construire un estimateur du paramètre θ ; les autres données contenues dans l'échantillon X sont superflues.

EXEMPLE 1. Soit $X \in \Pi_\lambda$. Montrons que la statistique $S = n\bar{\mathrm{x}} = \sum_{i=1}^{n} \mathrm{x}_i$ est exhaustive pour le paramètre λ de la loi de Poisson. Il nous faut montrer que la distribution de la position du point X sur la surface $\sum_{i=1}^{n} x_i = s$ (s est un entier) ne dépend pas de λ. Comme $\mathbf{P}(X = \mathbf{x}, \sum \mathrm{x}_i = s) = \mathbf{P}(X = \mathbf{x})$ pour $\sum_{i=1}^{n} x_i = s$, il vient

$$\mathbf{P}(X = \mathbf{x}|n\bar{\mathrm{x}} = s) = \begin{cases} \dfrac{\mathbf{P}(\mathrm{x}_1 = x_1, \ldots, \mathrm{x}_n = x_n)}{\mathbf{P}(n\bar{\mathrm{x}} = s)} & \text{si } \sum_{i=1}^{n} x_i = s, \\ 0 \quad \text{sinon.} \end{cases} \tag{1}$$

Puisque les x_i sont indépendants et $\sum_{i=1}^{n} x_i \in \Pi_{n\lambda}$, le second membre de (1) est égal à

$$\left(e^{-n\lambda} \frac{(n\lambda)^s}{s!}\right)^{-1} \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = \frac{s!}{n^s \prod_{i=1}^{n} x_i!}.$$

Donc, la distribution conditionnelle de X sachant que $S = s$ est confondue avec la distribution polynomiale $\mathbf{B}_p^s$ (cf. § 2) à n issues équiprobables (c'est-à-dire de probabilité $p = (1/n, \ldots, 1/n)$) et à s épreuves indépendantes. Il est évident que cette distribution ne dépend pas de λ, si bien que $S = n\bar{x}$ est une statistique exhaustive pour λ.

La notion de statistique exhaustive a été introduite par Fisher en 1922. Le théorème suivant de Neyman-Fisher, appelé *théorème de factorisation*, établit un critère simple d'existence d'une statistique exhaustive.

Supposons qu'est remplie la condition (A_μ) d'existence de la densité $f_\theta(x) = \frac{d\mathbf{P}_\theta}{d\mu}(x)$.

THÉORÈME 1. *La condition nécessaire et suffisante pour qu'une statistique S soit exhaustive pour θ est que la fonction de vraisemblance* $f_\theta(x) = \prod_{i=1}^{n} f_\theta(x_i)$ *se représente sous la forme*

$$f_\theta(x) = \psi(S(x), \theta)h(x) \quad [\mu^n]\text{-presque partout,} \tag{2}$$

où chacune des fonctions $\psi \geqslant 0$ *et* $h \geqslant 0$ *dépend uniquement de ses arguments,* $\psi(s, \theta)$ *est mesurable par rapport à s, et* $h(x)$ *mesurable par rapport à* x.

Il est clair que la représentation (2) n'est pas unique. Ses composantes sont définies à une fonction strictement positive de $S(x)$ près.

Dans l'exemple ci-dessus relatif à la distribution de Poisson on a

$$f_\lambda(x) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{n\bar{x}} \prod_{i=1}^{n} \frac{1}{x_i!},$$

$$n\bar{x} = \sum_{i=1}^{n} x_i,$$

de sorte que pour $S = n\bar{x}$ on peut poser

$$\psi(S, \lambda) = e^{-n\lambda} \lambda^S h, \quad h(x) = \sum_{i=1}^{n} \frac{1}{x_i!}.$$

De là il s'ensuivra en vertu du théorème 1 que $S = n\bar{x}$ est une statistique exhaustive.

Nous produirons la démonstration du théorème 1 pour deux cas seulement : le cas discret et le cas « régulier ». Dans le cas général, cette démonstration est accessible dans l'Annexe IV.

Dans le *cas discret,* μ est une mesure cardinale sur l'ensemble dénombrable $\mathscr{X}$ des valeurs possibles de x_1 et par suite $f_\theta(x) = \mathbf{P}_\theta(x_1 = x)$, $x \in \mathscr{X}$. Supposons tout d'abord que (2) est réalisée. Pour tout point $x \in \mathscr{X}^n$ fixe, on a alors

$$\mathbf{P}_\theta(X = x \mid S(X) = S(x)) = \frac{\mathbf{P}_\theta(X = x,\ S(X) = S(x))}{\mathbf{P}_\theta(S(X) = S(x))}. \qquad (3)$$

Comme $\{X = x,\ S(X) = S(x)\} = \{X = x\}$, le second membre de (3) vaut

$$\frac{\mathbf{P}_\theta(X = x)}{\mathbf{P}_\theta(S(X) = S(x))} = \frac{f_\theta(x)}{\sum\limits_{y : S(y) = S(x)} f_\theta(y)} =$$

$$= \frac{\psi(S(x), \theta)h(x)}{\sum\limits_{y : S(y) = S(x)} \psi(S(x), \theta)h(y)} = \frac{h(x)}{\sum\limits_{y : S(y) = S(x)} h(y)}.$$

Donc, $\mathbf{P}_\theta(X = x \mid S(X) = S(x))$ ne dépend pas de θ.

Réciproquement, si le premier membre de (3) est indépendant de θ, en le désignant par $h(x)$, on déduit de (3)

$$\mathbf{P}_\theta(X = x) = f_\theta(x) = \mathbf{P}_\theta(X = x\ ;\ S(X) = S(x)) = h(x)\mathbf{P}_\theta(S(X) = S(x)),$$

où $\mathbf{P}_\theta(S(X) = S(x)) = \psi(S(x), \theta)$ ne dépend que de $S(x)$ et de θ. ◄

La démonstration du théorème 1 est légèrement plus compliquée dans le cas « régulier » où μ est la mesure de Lebesgue dans R et la statistique $S(X)$, une fonction régulière de X telle qu'existe un changement de variables $y_1 = S(x)$, $y_2 = y_2(x), \ldots, y_n = y_n(x)$ tel que $x_i = x_i(y_1, \ldots, y_n)$ et $J = \left|\frac{\partial x_i}{\partial y_j}\right| \neq 0$. Du cours d'analyse classique sur le changement de variable sous le signe d'intégration, on sait que dans ce cas la densité de la variable aléatoire $Y = (S(X), y_2(X), \ldots, y_n(X))$ sera égale à

$$g_\theta(y) = f_\theta(x)|J|, \quad y = (y_1, \ldots, y_n).$$

La densité de la variable aléatoire $y_1(X) = S(X)$ vaudra

$$g_\theta^{(1)}(y_1) = \int_{R^{n-1}} g_\theta(y)dy_2 \ldots dy_n = \int_{R^{n-1}} f_\theta(x)|J|dy_2 \ldots dy_n,$$

quant à la densité conditionnelle Y sachant que $S(X) = s$, elle sera par conséquent définie par l'expression

$$\varphi(y|s) = \frac{g_\theta(y)}{g_\theta^{(1)}(s)} = \frac{f_\theta(x)|J|}{g_\theta^{(1)}(s)} \quad \text{pour} \quad y_1 = s.$$

Après ces remarques préliminaires la démonstration du théorème 1 pour le cas « régulier » s'effectue comme pour le cas discret. En effet, si (2) est réalisée, on a

$$\varphi(y|s) = \frac{\psi(s, \theta)h(x)|J|}{\int\limits_{R^{n-1}} \psi(s, \theta)h(x)|J|dy_2 \ldots dy_n}.$$

Dans cette relation on peut simplifier par $\psi(s, \theta)$. Ceci exprime que la distribution conditionnelle de Y, donc de X, par rapport à la condition $S(X) = s$ ne dépend pas de θ.

Réciproquement, si $\varphi(y|s)$ ne dépend pas de θ, il vient

$$f_\theta(x) = \frac{\varphi(y|s)g_\theta^{(1)}(s)}{|J|} \quad \text{pour} \quad s = S(x).$$

Ceci signifie que (2) est réalisée pour $\psi(s, \theta) = g_\theta^{(1)}(s)$, $h(x) = \varphi(y|s)/|J|$. ◄

EXEMPLE 2. Soit $X \in \Phi_{\alpha, \sigma^2}$. Le paramètre $\theta = (\alpha, \sigma^2)$ est à deux dimensions. On a

$$f_\theta(X) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\alpha)^2}{2\sigma^2}} = \sigma^{-n}(2\pi)^{-n/2}\exp\left\{-\frac{\sum(x_i-\alpha)^2}{2\sigma^2}\right\} =$$

$$= \sigma^{-n}\exp\left\{-\frac{\sum x_i^2 - 2\alpha n\bar{x} + n\alpha^2}{2\sigma^2}\right\}(2\pi)^{-n/2}.$$

En posant $S = (S_1, S_2)$, $S_1 = n\bar{x}$, $S_2 = \sum\limits_{i=1}^{n} x_i^2$, on obtient la représentation (2), où

$$\psi(S, \theta) = \sigma^{-n}\exp\left\{-\frac{S_2 - 2\alpha S_1 + \alpha^2}{2\sigma^2}\right\}, \quad h(X) = (2\pi)^{-n/2}.$$

On aurait pu certes rapporter le facteur $(2\pi)^{-n/2}$ à la fonction ψ en posant $h(X) = 1$.

On trouve donc que la statistique (S_1, S_2) est une statistique vectorielle exhaustive pour (α, σ^2). De toute l'information contenue dans l'échantillon, il nous suffit de connaître $\bar{x}$ et $\sum x_i^2$.

Nous proposons au lecteur de trouver les statistiques exhaustives pour toutes les familles de distrit tions citées dans le § 2.

Arrêtons-nous en détail sur l'une de ces familles.

EXEMPLE 3. Soit $X \in \mathbf{U}_{0,\,\theta}$. La condition (A_μ) est remplie pour la mesure de Lebesgue et

$$f_\theta(X) = \begin{cases} \theta^{-n} & \text{si } \mathrm{x}_i \in [0, \theta], \quad i = 1, \ldots, n, \\ 0 & \text{sinon.} \end{cases}$$

Soient $\mathrm{x}_{(1)} = \min \mathrm{x}_i$, $\mathrm{x}_{(n)} = \max \mathrm{x}_i$. Alors, comme dans l'exemple 6.5, la fonction $f_\theta(X)$ peut se mettre sous la forme $f_\theta(X) = \psi(\mathrm{x}_{(n)}, \theta)h(X)$, où

$$h(X) = \begin{cases} 1 & \text{si } \mathrm{x}_{(1)} \geqslant 0, \\ 0 & \text{sinon,} \end{cases}$$

$$\psi(s, \theta) = \begin{cases} \theta^{-n} & \text{si } s \leqslant \theta, \\ 0 & \text{sinon.} \end{cases}$$

Ce qui exprime que $S(X) = \mathrm{x}_{(n)}$ est une statistique exhaustive pour θ.

Le lecteur peut s'assurer de façon analogue que si $X \in \mathbf{U}_{\theta,\,1+\theta}$, la statistique $S(X) = (\mathrm{x}_{(1)}, \mathrm{x}_{(n)})$ est une statistique exhaustive pour le paramètre θ. On obtient la même statistique exhaustive pour le paramètre $\theta = (a, b)$ si $X \in \mathbf{U}_{a,\,b}$.

Voici deux corollaires du théorème 1.

COROLLAIRE 1. *Si S est une statistique exhaustive pour θ, l'estimateur du maximum de vraisemblance ne dépend que de S.*

Plus exactement, l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ ne dépend pas de X si $S(X)$ est fixe.

Ce corollaire est évident, puisque l'estimation par le maximum de vraisemblance est la valeur de θ qui réalise le maximum de $f_\theta(X) = \psi(S(X), \theta)h(X)$ ou, ce qui est équivalent, le maximum de $\psi(S(X), \theta)$.

COROLLAIRE 2. *Si S est une statistique exhaustive et φ une fonction telle que l'application $u = \varphi(v)$ est bijective et mesurable avec sa réciproque, alors $S_1 = \varphi(S)$ est aussi une statistique exhaustive.*

Ce corollaire est évident, lui aussi, puisque la fonction $\psi(S, \theta)$ de (2) peut être mise sous la forme $\psi(\varphi^{-1}(S_1), \theta) = \psi_1(S_1, \theta)$.

Nous avons encore le critère suivant d'exhaustivité d'une statistique S.

THÉORÈME 2. *Une condition nécessaire et suffisante pour qu'une statistique S soit exhaustive est que pour toute distribution a priori* $\mathbf{Q}$ *du paramètre θ la distribution a posteriori* $\mathbf{Q}_X$ *ne dépende de X que par l'intermédiaire de $S(X)$ (c'est-à-dire reste invariable sur la surface $S(X) = s$).*

DÉMONSTRATION. Soient S une statistique exhaustive, $q(t)$ la densité de $\mathbf{Q}$ par rapport à une mesure λ. La densité *a posteriori* $q(t \mid X)$ par rapport à cette mesure est donnée par la formule de Bayes

$$q(t \mid X) = \frac{f_t(X)q(t)}{\int f_u(X)q(u)\lambda(du)} = \frac{\psi(S(X), t)q(t)}{\int \psi(S(X), u)q(u)\lambda(du)}.$$

Prouvons maintenant la condition suffisante du théorème. Choisissons la distribution *a priori* de telle sorte que $q(t) > 0$ partout sur Θ et que l'on ait pour tous les t

$$f_t(X) = \frac{q(t \mid X)f(X)}{q(t)}, \quad f(X) = \int f_u(X)q(u)\lambda(du).$$

Si $q(t \mid X) = g(t, S(X))$, on obtient la représentation (2) en posant $\psi(s, t) = g(t, s)/q(t)$, $h(X) = f(X)$. ◄

COROLLAIRE 3. *Si S est une statistique exhaustive, tous les estimateurs bayésiens et tous les estimateurs minimax définis à l'aide du théorème* 11.2 *ne dépendent que de S.*

Dans la suite nous aurons à maintes reprises la confirmation que la statistique exhaustive S contient une information exhaustive sur θ.

§ 13*. Statistiques exhaustives minimales

Penchons-nous maintenant sur le choix des statistiques exhaustives. Il est clair qu'il en existe beaucoup. Par exemple, la statistique $S(X) \equiv X$ est visiblement toujours exhaustive. On l'appelle *statistique exhaustive triviale.* Mais l'on s'intéressera (et l'on verra pourquoi dans la suite) aux statistiques plus « économiques ». Il s'avère qu'il n'est pas toujours possible de construire des statistiques exhaustives qui soient sensiblement plus « économiques » que la statistique exhaustive triviale. On reviendra sur cette question une fois qu'on aura défini rigoureusement les notions de statistiques exhaustives « économiques ». A cet effet, munissons l'ensemble des statistiques exhaustives (pour un paramètre θ) d'une relation d'ordre partiel.

DÉFINITION 1. On dira qu'une statistique S_1 est *subordonnée* à une statistique S_2 si S_1 est une fonction mesurable de S_2 : $S_1 = \varphi(S_2)$.

Cette relation exprime précisément que S_1 est plus « économique » que S_2.

DÉFINITION 2. On dit que des statistiques S_1 et S_2 sont *équivalentes* si S_1 est subordonnée à S_2, et S_2 à S_1.

Il est évident que S_1 est équivalente à S_2 si et seulement si $S_1 = \varphi(S_2)$ et φ est une application bijective mesurable avec sa réciproque.

DÉFINITION 3. On dit qu'une statistique exhaustive S_0 est *minimale* si elle est subordonnée à toute autre statistique exhaustive S.

Les statistiques exhaustives minimales sont les plus économiques. Si l'on a réussi à construire une statistique exhaustive minimale S, il est impossible de réduire les données tout en conservant l'exhaustivité. Les autres données contenues dans l'échantillon peuvent être considérées comme engendrées par un mécanisme aléatoire indépendant de θ. Elles ne recèlent aucune information sur θ.

Les notions introduites ainsi que la notion initiale de statistique exhaustive peuvent être exposées sous une forme légèrement généralisée dans le langage des tribus, langage qui, dans bien des cas, est plus commode et plus suggestif. Tout au début — dans la définition 1 du paragraphe précédent — on peut remplacer la distribution conditionnelle $\mathbf{P}_\theta(X \in B | S)$ par la distribution conditionnelle $\mathbf{P}_\theta(X \in B | \mathfrak{A})$ par rapport à une sous-tribu $\mathfrak{A} \subset \mathfrak{B}^n_{\mathscr{X}}$ et appeler $\mathfrak{A}$ *tribu exhaustive* s'il existe une distribution $\mathbf{P}_\theta(X \in B | \mathfrak{A})$ indépendante de θ.

Le théorème de factorisation reste en vigueur si la fonction $\psi(S(X), \theta)$ est remplacée par une fonction $\psi(X, \theta)$ $\mathfrak{A}$-mesurable par rapport à X. La démonstration de ce théorème, qui est insérée dans l'Annexe IV, reste pratiquement la même.

On peut maintenant définir une statistique exhaustive S comme une statistique pour laquelle la tribu $\sigma(S)$ qu'elle engendre est exhaustive.

La relation « être subordonnée à » entre les statistiques exhaustives (cf. définition 1), traduite dans le langage des tribus, n'implique l'introduction d'aucune notion supplémentaire et est confondue avec l'immersion des tribus : S_1 est subordonnée à S_2 si $\sigma(S_1) \subset \sigma(S_2)$. Donc, S_1 est plus économique que S_2 si la tribu $\sigma(S_1)$ est plus pauvre (plus grossière) que $\sigma(S_2)$. L'équivalence de S_1 et de S_2 exprime que $\sigma(S_1) = \sigma(S_2)$.

La *tribu exhaustive minimale* $\mathfrak{A}_0$ se définit comme une tribu qui s'immerge dans toute tribu exhaustive.

Il existe toujours une tribu exhaustive minimale. Pour s'en assurer, on remarquera préalablement qu'en vertu du théorème 2 de l'Annexe IV il existe une distribution (discrète) $\mathbf{Q}$ sur Θ telle que toutes les $\mathbf{P}_\theta$ sont absolument continues par rapport à la distribution $\mathbf{P}_\mathbf{Q} = \int \mathbf{P}_t \mathbf{Q}(dt)$.

Ceci exprime que soit $f_\mathbf{Q}(X) = \int f_t(X)\mathbf{Q}(dt) > 0$ pour tous les X, soit l'égalité $f_\mathbf{Q}(X) = 0$ entraîne que $f_\theta(X) = 0$ pour tout θ. On dit alors que $\mathbf{P}_\mathbf{Q}$ *domine* la famille $\{\mathbf{P}_\theta\}$, si bien qu'on aurait pu prendre $\mathbf{P}_\mathbf{Q}$ pour mesure μ. La densité de la distribution $\mathbf{P}_\theta$ par rapport à cette mesure est égale à

$$\frac{d\mathbf{P}_\theta}{d\mathbf{P}_\mathbf{Q}}(x) = \frac{f_\theta(x)}{f_\mathbf{Q}(x)} \equiv r(x, \theta).$$

Il est clair (comparer avec le théorème 12.2) que si S est une statistique exhaustive, $r(x, \theta)$ ne dépend de x que par l'intermédiaire de $S(x)$.

THÉORÈME 1. *La tribu* $\mathfrak{A}_0 = \sigma(r(X, \theta)\,; \theta \in \Theta)$ *engendrée par les variables aléatoires* $r(X, \theta) = f_\theta(X)/f_{\mathbf{Q}}(X)$ *pour diverses valeurs de* $\theta \in \Theta$, *est une tribu exhaustive minimale.*

DÉMONSTRATION. Elle est élémentaire. L'exhaustivité de $\mathfrak{A}_0$ résulte du théorème de factorisation et du fait que

$$f_\theta(X) = r(X, \theta) f_{\mathbf{Q}}(X), \tag{1}$$

où $f_{\mathbf{Q}}(X)$ ne dépend pas de θ et $r(X, \theta)$ est mesurable par rapport à $\mathfrak{A}_0$.

Supposons maintenant que $\mathfrak{A}$ est une tribu exhaustive quelconque. Alors $f_\theta(X) = \psi(X, \theta)h(X)$, où la fonction $\psi(X, \theta)$ est $\mathfrak{A}$-mesurable. Considérons la tribu $\mathfrak{A}_\psi \equiv \sigma(\psi(X, \theta), \theta \in \Theta) \subset \mathfrak{A}$. De la définition de $r(X, \theta)$ il résulte que

$$r(X, \theta) = \frac{\psi(X, \theta)}{\int \psi(X, t)\mathbf{Q}(dt)},$$

donc $\mathfrak{A}_0 \subset \mathfrak{A}_\psi \subset \mathfrak{A}$. ◄

A ce théorème et au théorème 12.2 est liée une autre proposition utile. Considérons l'approche bayésienne du problème où θ est une variable aléatoire de distribution *a priori* $\mathbf{Q}$. Supposons que $q(t) > 0$ est la densité de cette distribution par rapport à une mesure convenable λ sur Θ. La densité *a posteriori* sera alors égale à

$$q(t \mid X) = \frac{f_t(X)q(t)}{f_{\mathbf{Q}}(X)} = r(X, t)q(t),$$

donc la tribu exhaustive minimale $\mathfrak{A}_0$ peut être considérée comme engendrée par la distribution *a posteriori*, soit :

$$\mathfrak{A}_0 = \sigma(q(t \mid X)\ ;\ t \in \Theta).$$

La détermination des distributions $\mathbf{Q}$ et $\mathbf{P}_{\mathbf{Q}}$ figurant dans le théorème 1 ne pose aucun problème. Si, par exemple, le support $N_{\mathbf{P}_\theta}$ de la distribution $\mathbf{P}_\theta$ ne dépend pas de θ, ce qui est le cas de la plupart des distributions mentionnées dans le § 2, on peut prendre $\mathbf{P}_{\mathbf{Q}} = \mathbf{P}_{\theta_0}$ pour tout $\theta_0 \in \Theta$.

Nous disposons ainsi d'un théorème d'existence et d'une méthode de construction des tribus exhaustives minimales *).

---

*) On peut établir l'existence d'une tribu exhaustive minimale $\mathfrak{A}_0$ d'une autre manière en prouvant qu'elle est l'intersection de toutes les tribus exhaustives complétées.

Cependant dans la plupart des cas il nous sera plus commode de manipuler les statistiques. Le principal objectif de ce paragraphe est la recherche des statistiques exhaustives minimales.

Mais tout d'abord comment peut-on s'assurer qu'une statistique exhaustive S_0 est minimale ?

Un moyen consiste à utiliser le théorème 1. Si $\sigma(S_0)$ est confondue avec la tribu engendrée par $f_\theta(X)/f_Q(X)$, alors S_0 est une statistique exhaustive minimale.

EXEMPLE 1. Nous avons vu que la statistique $S = n\bar{x}$ est exhaustive pour le paramètre λ de la distribution de Poisson Π_λ. C'est une statistique exhaustive minimale, puisque $\sigma(S)$ est confondue de toute évidence avec la tribu engendrée par $f_\lambda(X)/f_{\lambda_1}(X) = e^{n(\lambda_1 - \lambda)}(\lambda/\lambda_1)^S$ (nous avons envisagé ici une distribution Q concentrée au point λ_1).

EXEMPLE 2. Soit $X \in U_{0,\theta}$. La statistique $S = x_{(n)} = \max x_i$ est alors une statistique exhaustive minimale. En effet, prenons pour Q une distribution quelconque sur $[0, \infty[$, de densité $q(t) > 0$ pour tout $t > 0$. Alors

$$f_\theta(X) = \begin{cases} \theta^{-n}, & \theta \geqslant S, \\ 0, & \theta < S, \end{cases}$$

$$f_Q(X) = \int_0^\infty f_t(X)q(t)dt = \int_S^\infty t^{-n}q(t)dt > 0$$

pour tout X. De plus, $S = \sup\{\theta : f_\theta(X)/f_Q(X) = 0\}$. Ceci exprime que S est mesurable par rapport à la plus petite tribu $\mathfrak{A}_0$, $\sigma(S) \subset \mathfrak{A}_0$ et par suite, S est une statistique exhaustive minimale.

Il existe un autre procédé de détermination des statistiques exhaustives minimales qui est également lié à la fonction de vraisemblance. En effet, toute statistique, et en particulier toute statistique exhaustive engendre une partition de l'espace des échantillons en classes d'équivalence, c'est-à-dire en sous-ensembles de points x en lesquels $S(x)$ prend la même valeur.

Si S_1 est subordonnée à S_2, i.e. $S_1 = \varphi(S_2)$, il est évident que la partition pour S_1 sera moins fine puisque les classes d'équivalence pour S_2 sont contenues dans les classes d'équivalence pour S_1. Donc, à une statistique exhaustive minimale est associée la plus « grosse » de toutes les partitions engendrées par les statistiques exhaustives.

On peut envisager simplement des partitions de l'espace en classes d'équivalence sans les relier directement aux statistiques. Désignons par $D(x)$ la classe d'équivalence contenant le point x. Chaque classe est définie de façon unique par l'un quelconque de ses points. *On dira qu'une partition en classes D est exhaustive si*

$$f_\theta(x) = \varphi(x, \theta)h(x), \tag{2}$$

où $\varphi(x, \theta) = \varphi(x_0, \theta)$ *est constante pour* $x \in D(x_0)$ (i.e. $\varphi(x, \theta) = \text{const}$ à l'intérieur de la classe d'équivalence). Si les classes $D(x)$ sont définies par les relations $S(x) = s$, il découle alors immédiatement du théorème 11.1 que la statistique $S(x)$ est exhaustive si et seulement s'il en est de même de la partition en classes D.

Considérons maintenant la partition suivante : prenons un point x_0 et décrétons que x appartient à la classe $D(x_0)$ si le rapport

$$\frac{f_\theta(x)}{f_\theta(x_0)} = h(x, x_0) \tag{3}$$

est indépendant de θ. Il est évident que $D(x_1) = D(x_2) = D(x_0)$ si $x_1 \in D(x_0)$ et $x_2 \in D(x_0)$, de sorte que la règle (3) engendre une partition de l'espace tout entier en classes disjointes.

Cette partition correspond à celle engendrée par une statistique exhaustive minimale S.

En effet, soit S une statistique exhaustive minimale. Prenons un point x_0 quelconque. Sur la surface $S(x) = S(x_0)$, le rapport $f_\theta(x)/f_\theta(x_0)$ est égal à $h(x)/h(x_0)$ et ne dépend donc pas de θ. Par conséquent, la partition en classes D est au moins aussi grosse que la partition pour S.

D'autre part, cette partition est exhaustive. En effet, à toute surface D on peut associer l'un quelconque de ses points x_D qui la définira de façon unique. Considérons la fonction $x_0(x)$ définie par la relation $x_0(x) = x_D$ si $x \in D$. Alors, en vertu de (3), pour $x \in D$, on a

$$f_\theta(x) = f_\theta(x_D)h(x, x_D) = f_\theta(x_0(x))h(x, x_0(x)), \tag{4}$$

ce qui exprime que (2) est réalisée.

Les considérations précédentes n'étaient pas rigoureuses du tout, car elles n'étaient pas liées à la mesurabilité des fonctions figurant dans (4).

Ce qui précède peut être résumé comme suit. *Soit donnée une statistique $S(X)$ telle que $S(x) = S(x_0)$ si et seulement si le rapport* (3) *est indépendant de θ. Dans ce cas, S est une statistique exhaustive minimale.*

Contrairement aux approches liées au théorème 1, qui considéraient les rapports $f_\theta(x)/f_Q(x)$ ou $f_\theta(x)/f_{\theta_1}(x)$ pour des θ et θ_1 différents (ces rapports sont souvent appelés rapports de vraisemblance), la règle formulée ci-dessus utilise le rapport $f_\theta(x)/f_\theta(x_0)$ pour les mêmes valeurs du paramètre θ. Ainsi, dans l'exemple 1, le rapport

$$f_\lambda(x)/f_\lambda(x_0) = \Pi\lambda^{x_i - x_{i0}} x_{i0}!/x_i! = \lambda^{n(\bar{x} - \bar{x}_0)}\Pi x_{i0}!/x_i!$$

sera indépendant de λ si et seulement si $\bar{x} = \bar{x}_0 = \frac{1}{n}\sum_{i=1}^{n} x_{i0}$, où x_{i0} sont les coordonnées du vecteur x_0. Ceci suffit pour conclure que $S(x) = \bar{x}$ est une statistique exhaustive minimale.

Appliquons maintenant la règle proposée à l'étude d'un exemple où il n'existe pas de statistiques exhaustives « économiques ». Remarquons tout d'abord que l'échantillon ordonné $S_V = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ associé à l'échantillon X est visiblement toujours une statistique exhaustive, puisque $f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{k=1}^{n} f_\theta(x_{(k)})$. Cette statistique est un « peu plus économique » que l'échantillon X. De là il s'ensuit en particulier que toute statistique exhaustive minimale est invariante par une permutation des coordonnées x_i de l'échantillon X.

Si la densité $f_\theta(x)$ est symétrique, i.e. $f_\theta(-x) = f_\theta(x)$ pour tous les θ, il est évident qu'il existera alors une statistique exhaustive un « peu plus économique », notée S_V^2, qui est constituée de l'ensemble $(x_1^2, \ldots, x_n^2)$ rangé dans l'ordre de grandeur croissante.

EXEMPLE 3. Si $X \in \mathbf{K}_{0,\sigma}$, i.e. x_i admettent une distribution de Cauchy de paramètre $\theta = \sigma$ et de densité

$$k_{0,\sigma}(x) = \frac{\sigma}{\pi(x^2 + \sigma^2)},$$

la statistique S_{ν^2} est une statistique exhaustive minimale. En effet, dans ce cas

$$f_\sigma(x) = \left(\frac{\sigma}{\pi}\right)^n \prod_{i=1}^{n} (x_i^2 + \sigma^2)^{-1},$$

de sorte que

$$\frac{f_\sigma(x)}{f_\sigma(x_0)} = \prod_{i=1}^{n} \frac{x_{i0}^2 + \sigma^2}{x_i^2 + \sigma^2} \tag{5}$$

est le rapport de deux polynômes de σ^2. Ce rapport est indépendant de σ si et seulement si les coefficients des puissances respectives de σ^2 du numérateur et du dénominateur sont confondus. Ce qui a lieu si et seulement si les ensembles des « zéros » $\{-x_{i0}^2\}$ et $\{-x_i^2\}$ sont confondus. En d'autres termes, une condition nécessaire et suffisante pour que le rapport (5) soit indépendant de σ est que le point $x^2 = (x_1^2, \ldots, x_n^2)$ admette les mêmes coordonnées que le point x_0^2 à une permutation près. Ceci exprime que S_{ν^2} est une statistique exhaustive minimale.

On démontre de façon analogue que S_ν est une statistique exhaustive minimale pour le paramètre α et, par suite, pour le paramètre $\theta = (\alpha, \sigma)$ de la distribution $\mathbf{K}_{\alpha,\sigma}$.

On obtient un autre exemple dans lequel S_{ν} est une statistique exhaustive minimale en considérant la famille

$$\mathbf{P}_{\alpha,\theta_1,\theta_2} = \alpha \mathbf{P}_{\theta_1} + (1 - \alpha)\mathbf{P}_{\theta_2}, \; \alpha \in [0, 1],$$

où $\{\mathbf{P}_\theta\}$ est une famille exponentielle (cf. § 15 ; pour $\mathbf{P}_\theta$ on peut prendre une distribution normale ou une distribution de Poisson) et l'un au moins des paramètres α, θ_1 ou θ_2 est inconnu.

Prouvons maintenant un théorème qui nous fournit une méthode élémentaire de construction des statistiques exhaustives minimales.

Pour simplifier l'exposé, on traitera le cas d'un paramètre θ scalaire.

THÉORÈME 2. *Supposons que la fonction de vraisemblance $f_\theta(x)$ considérée comme une fonction de θ est continue à droite (ou à gauche) pour tout x. Si l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est unique et est une statistique exhaustive, alors il est une statistique exhaustive minimale.*

DÉMONSTRATION. Soit S une statistique exhaustive quelconque. Le théorème sera démontré quand on aura établi que $\hat{\theta}^*$ est mesurable par rapport à $\sigma(S)$ et par suite $\hat{\theta}^*$ est subordonné à S.

Le théorème de factorisation affirme que

$$f_\theta(x) = \psi(S(x), \theta)h(x) \quad [\mu^n]\text{-presque partout,} \tag{6}$$

où $h(x)$ est une fonction mesurable par rapport à x, $\psi(s, t)$ une fonction continue (à droite ou à gauche) par rapport à t et mesurable par rapport à s. Comme $\mathbf{P}_\theta$ ne varie pas si la densité $f_\theta(x)$ change sur un ensemble μ^n-négligeable, on peut admettre que (6) est vérifiée pour tous les x.

D'après (6), le point de maximum absolu de $f_\theta(x)$ est aussi point de

maximum absolu de $\psi(S(x), \theta)$. Puisque $\hat{\theta}^*$ est unique, il vient donc

$$\{\hat{\theta}^* < t\} = \{\sup_{\theta < t} \psi(S(X), \theta) > \sup_{\theta \geq t} \psi(S(X), \theta)\}.$$

Comme $\psi(S(X), \theta)$ est continue à droite (ou à gauche) par rapport à θ pour toute $S(X)$, il existe un ensemble dénombrable partout dense $\Theta_d = \{\theta_j\}_{j=1}^{\infty} \subset \Theta$ (le même pour toutes les $S(X)$) tel que

$$\sup_{\theta < t} \psi(S(X), \theta) = \sup_{\substack{\theta_j < t \\ \theta_j \in \Theta_d}} \psi(S(X), \theta_j). \tag{7}$$

Cette relation sera valable aussi pour le domaine $\theta \geqslant t$. Comme $\psi(S(X), \theta_j)$ sont mesurables par rapport à $\sigma(S)$, les valeurs $\sup_{\theta < t} \psi(S, \theta)$ et $\sup_{\theta \geq t} \psi(S, \theta)$ seront, en vertu de (7), des variables aléatoires qui seront mesurables aussi par rapport à $\sigma(S)$. Donc, $\{\hat{\theta}^* < t\} \in \sigma(S)$ et le théorème est prouvé. ◄

L'exhaustivité de l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est essentielle dans cette proposition, puisque $\hat{\theta}^*$ n'est pas tenu de l'être. On obtient sans peine un exemple illustrant cette situation en considérant une famille quelconque de distributions $\{\mathbf{P}_\theta\}$ avec un paramètre θ scalaire et une statistique exhaustive minimale vectorielle S. Dans ce cas, l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ sera scalaire aussi, de sorte que la tribu $\sigma(S)$ sera plus riche que $\sigma(\hat{\theta}^*)$ et par suite l'inclusion $\sigma(S) \subset \sigma(\hat{\theta}^*)$, qui découle de la minimalité de S et de l'exhaustivité de $\hat{\theta}^*$, est impossible.

EXEMPLE 4. Soit $X \in \mathbf{U}_{\theta, 1+\theta}$, $\Theta = R$. Comme dans l'exemple 6.4, on a

$$f_\theta(X) = \begin{cases} 1 & \text{si} \quad \theta \leqslant x_{(1)} \leqslant x_{(n)} \leqslant 1 + \theta, \\ 0 & \text{sinon,} \end{cases}$$

de sorte que $f_\theta(X)$ dépend de X par l'intermédiaire seulement de $x_{(1)}$ et de $x_{(n)}$. Ceci exprime que $S = (x_{(1)}, x_{(n)})$ est une statistique exhaustive. Prise séparément, aucune des quantités $x_{(1)}$ et $x_{(n)}$ n'est une statistique exhaustive, la preuve étant donnée par les relations suivantes :

$$\mathbf{P}(x_{(1)} \geqslant u, x_{(n)} < v) = \prod_{i=1}^{n} \mathbf{P}(x_i \in [u, v[) =$$
$$= (v - u)^n \quad \text{pour} \quad u \geqslant \theta, \quad v \leqslant 1 + \theta, \quad v > u.$$

Donc, la densité conjointe de la distribution de $(x_{(1)}, x_{(n)})$ sera égale à

$$g(u, v) = \begin{cases} n(n-1)(v-u)^{n-2} & \text{si } u \geqslant \theta, v \leqslant 1 + \theta, v > u, \\ 0 & \text{sinon.} \end{cases}$$

D'autre part, $\mathbf{P}(\mathrm{x}_{(1)} \geqslant u) = (1 + \theta - u)^n$ pour $\theta \leqslant u \leqslant 1 + \theta$, de sorte que la densité de $\mathrm{x}_{(1)}$ est

$$g(u) = n(1 + \theta - u)^{n-1} \quad \text{pour} \quad \theta \leqslant u \leqslant 1 + \theta.$$

De là on déduit sans peine que la densité conditionnelle $g(v \mid u)$ de la variable $\mathrm{x}_{(n)}$ sachant que $\mathrm{x}_{(1)} = u$ (donc la distribution conditionnelle correspondante) dépendra de θ. Ceci exprime que $\mathrm{x}_{(1)}$ (de même que $\mathrm{x}_{(n)}$) n'est pas une statistique exhaustive. Comme on peut adopter $\hat{\theta}^* = \mathrm{x}_{(1)}$ en qualité d'estimateur du maximum de vraisemblance $\hat{\theta}^*$ (cf. exemple 6.4), cela démontre que $\hat{\theta}^*$ n'est pas une statistique exhaustive pour la famille $\mathbf{U}_{\theta,\, 1+\theta}$.

Nous proposons au lecteur de s'assurer à l'aide du théorème 1 que $S = (\mathrm{x}_{(1)}, \mathrm{x}_{(n)})$ est une statistique exhaustive minimale pour $\mathbf{U}_{\theta,\, 1+\theta}$.

La condition d'exhaustivité de $\hat{\theta}^*$ du théorème 2 sera automatiquement remplie si l'on admet qu'il existe une statistique exhaustive scalaire S_0 pour laquelle la fonction φ de l'égalité $\hat{\theta}^* = \varphi(S_0)$ sera biunivoque (c'est-à-dire que $\hat{\theta}^*$ et S_0 seront équivalentes).

§ 14. Construction des estimateurs efficaces à partir des statistiques exhaustives. Statistiques complètes

DÉFINITION 1. Un estimateur θ^* est *exhaustif* s'il est une statistique exhaustive.

1. Cas scalaire. On admettra que θ est un paramètre scalaire. Soit K_b la classe des estimateurs biaisés θ^* de biais $b(\theta)$, c'est-à-dire que $\theta^* \in K_b$ si $a(\theta) = \mathbf{E}_\theta \theta^* = \theta + b(\theta)$. Pour $\theta^* \in K_b$, on a

$$\mathbf{E}_\theta(\theta^* - \theta)^2 = \mathbf{E}_\theta(\theta^* - a(\theta))^2 + (a(\theta) - \theta)^2 = \mathbf{V}_\theta \theta^* + b^2(\theta).$$

On omettra parfois l'indice θ des symboles $\mathbf{E}_\theta$ et $\mathbf{V}_\theta$ dans ce paragraphe.

La proposition suivante a été établie indépendamment par Blackwell, Rao et Kolmogorov.

THÉORÈME 1. *Si S est une statistique exhaustive et $\theta^* \in K_b$, la fonction $\theta^*_S = \mathbf{E}_\theta(\theta^* \mid S)$ est un estimateur doué des propriétés suivantes :*

1) $\theta^*_S \in K_b$,

2) *θ^*_S dépend de X seulement par l'intermédiaire de $S(X)$,*

3) $\mathbf{E}_\theta(\theta^*_S - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta)^2$ *pour tout θ.*

La dernière relation se transforme en égalité si seulement $\theta^* = \theta^*_S$ presque partout par rapport à $\mathbf{P}_\theta$.

En d'autres termes, l'estimateur θ^* est uniformément amélioré si on lui applique l'opération $\mathbf{E}_\theta(\cdot \mid S)$ dans la classe K_b.

DÉMONSTRATION. Etant un estimateur, θ^*_S ne dépend pas de θ et est une fonction mesurable de X. Son indépendance par rapport à θ découle des propriétés des statistiques exhaustives, puisque la distribution de X est indépendante de θ pour S fixe (la quantité $\mathbf{E}_\theta(\theta^* | S)$ ne dépend généralement pas de θ pour S quelconque). D'autre part, d'après les propriétés de l'espérance mathématique conditionnelle, θ^*_S est une fonction mesurable de S, donc de X. Par conséquent, θ^*_S est un estimateur vérifiant la propriété 2) du théorème.

L'égalité

$$\mathbf{E}_\theta \theta^*_S = \mathbf{E}_\theta \mathbf{E}_\theta(\theta^* | S) = \mathbf{E}_\theta \theta^*,$$

qui prouve que $\theta^*_S \in K_b$, résulte aussi directement des propriétés de l'espérance mathématique conditionnelle. Par ailleurs,

$$\mathbf{E}_\theta(\theta^* - \theta)^2 = \mathbf{E}_\theta(\theta^* - \theta \pm \theta^*_S)^2 =$$
$$= \mathbf{E}_\theta(\theta^*_S - \theta)^2 + \mathbf{E}_\theta(\theta^* - \theta^*_S)^2 + 2\mathbf{E}_\theta(\theta^*_S - \theta)(\theta^* - \theta^*_S).$$

Les propriétés de l'espérance mathématique conditionnelle nous donnent encore

$$\mathbf{E}_\theta(\theta^*_S - \theta)(\theta^* - \theta^*_S) = \mathbf{E}_\theta \mathbf{E}_\theta[(\theta^*_S - \theta)(\theta^* - \theta^*_S) | S] =$$
$$= \mathbf{E}_\theta[(\theta^*_S - \theta)\mathbf{E}_\theta(\theta^* - \theta^*_S | S)] = 0,$$

donc

$$\mathbf{E}_\theta(\theta^* - \theta)^2 = \mathbf{E}_\theta(\theta^*_S - \theta)^2 + \mathbf{E}_\theta(\theta^* - \theta^*_S)^2. ◄$$

En fait, on aurait pu établir l'inégalité 3) du théorème 1 directement à partir de la propriété suivante : $(\mathbf{E}(\xi | S))^2 \leqslant \mathbf{E}(\xi^2 | S)$ de l'espérance mathématique conditionnelle, puisque alors

$$(\theta^*_S - \theta)^2 = [\mathbf{E}_\theta(\theta^* - \theta) | S)]^2 \leqslant \mathbf{E}_\theta[(\theta^* - \theta)^2 | S],$$
$$\mathbf{E}_\theta(\theta^*_S - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta)^2.$$

La proposition du théorème 1 admet l'interprétation suivante. Si S et T sont des statistiques exhaustives, $\theta^* = \varphi(T)$ et S est subordonnée à T, alors $\mathbf{E}_\theta(\theta^*_S - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^* - \theta)^2$.

En d'autres termes, plus la statistique exhaustive S est « économique » (ou plus la tribu correspondante est pauvre), plus les estimateurs θ^*_S sont meilleurs. Pour construire des estimateurs optimaux nous devons donc chercher des statistiques exhaustives *minimales* (ou les plus petites tribus). Ceci étant, les estimateurs de départ θ^* peuvent être de « mauvais » estimateurs qui par exemple ne sont même pas convergents. A cet égard, l'exemple suivant est instructif.

EXEMPLE 1. Soit $X \in \Pi_\lambda$. L'estimateur $\lambda^* = x_1$ est visiblement sans biais ($E\lambda^* = Ex_1 = \lambda$, $b(\lambda) = 0$) et n'est pas convergent, puisqu'il ne dépend pas de n. Une statistique exhaustive minimale pour λ est la statistique $S = n\bar{x} = \sum x_i$. De l'exemple 12.1 il s'ensuit que la distribution conditionnelle de x_1 par rapport à S est la distribution $B^S_{1/n}$:

$$P(x_1 = k | S = s) = C_s^k \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{s-k}.$$

Donc

$$\lambda^*_S = E(x_1 | S) = \sum_{k=1}^{S} k C_S^k \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{S-k} = \frac{S}{n} = \bar{x}.$$

Dans un exemple ultérieur on montrera que $\bar{x}$ est un estimateur efficace.

2. Cas vectoriel. Etablissons maintenant les analogues du théorème 1 pour le cas où θ et θ^* sont des vecteurs de R^k.

Comme en dimension un, le vecteur $b(\theta) = E_\theta\theta^* - \theta$ sera appelé *biais* de l'estimateur θ^* et la classe des estimateurs de biais b, désignée par K_b.

THÉORÈME 1A. *Soient S une statistique exhaustive et $\theta^* \in K_b$. L'estimateur $\theta^*_S = E_\theta(\theta^* | S)$ jouit alors des propriétés suivantes :*

1) $\theta^*_S \in K_b$,

2) *θ^*_S dépend uniquement de $S(X)$,*

3) *la dispersion quadratique moyenne de θ^*_S est inférieure à celle de θ^*, ou ce qui est équivalent, pour tout vecteur $a \in R^k$, on a*

$$E_\theta(\theta^*_S - \theta, a)^2 \leqslant E_\theta(\theta^* - \theta, a)^2. \qquad (1)$$

L'égalité (pour tout a) n'est possible que dans le cas où $\theta^* = \theta^*_S$ presque partout par rapport à P_θ.

DÉMONSTRATION. Les deux premières propositions sont évidentes. Les inégalités (1) résultent du théorème 1, puisque tout se ramène à l'étude d'estimateurs scalaires (θ^*, a) du paramètre (θ, a) et $E_\theta[(\theta^*, a) | S] = (\theta^*_S, a)$. Si l'égalité est réalisée dans (1) pour tout a, on aura alors $(\theta^*_S, a) = (\theta^*, a)$ presque partout. Ce qui signifie que $\theta^*_S = \theta^*$ presque partout. ◄

Dans le cas vectoriel les statistiques exhaustives jouent donc le même rôle que dans le cas scalaire : la forme quadratique $\sum \sigma_{ij} a_i a_j$, où $\sigma^2 = \|\sigma_{ij}\|$ est la matrice des moments d'ordre deux pour $\theta^*_S - \theta$, sera d'autant plus petite que le sera la tribu $\sigma(S)$ engendrée par S.

3. Statistiques complètes et estimateurs efficaces. Nous allons introduire maintenant un critère assez simple, basé sur la notion de complétude d'une statistique S, qui détermine l'impossibilité d'améliorer les estimateurs. Dési-

gnons la dimension de la statistique S par l. Généralement, $l \geqslant k$, où k est la dimension du paramètre θ.

Etant donné deux fonctions mesurables $f_1(s)$ et $f_2(s)$ de R^l, dans R^k, on écrira $f_1(s) = f_2(s)$ $[\mathscr{P}]$-presque partout, où $\mathscr{P}$ est une famille de distributions dans $(R^l, \mathfrak{B}^l)$, si $f_1(s) = f_2(s)$ partout sauf pour un ensemble N tel que $\mathbf{P}(N) = 0$, $\forall \mathbf{P} \in \mathscr{P}$.

DÉFINITION 2. On dira qu'une famille de distributions $\mathscr{G} = \{\mathbf{G}_\theta\}$ dans $(R^l, \mathfrak{B}^l)$, dépendant d'un paramètre à k dimensions $\theta \in \Theta \subset R^k$ est *complète* si l'égalité

$$\int y(s)\mathbf{G}_\theta(ds) = 0 \quad \text{pour tout} \quad \theta \in \Theta \tag{2}$$

entraîne $y(s) = 0$ $[\mathscr{G}]$-presque partout. L'équation (2) est envisagée dans la classe des fonctions $y : R^l \to R^k$ pour lesquelles existe l'intégrale (2).

DÉFINITION 3. Une statistique S est *complète* si la famille $\mathscr{G}$ de ses distributions $\mathbf{G}_\theta$ induites par une distribution $\mathbf{P}_\theta$ dans $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ est complète.

Pour les statistiques, l'équation (2) peut être mise sous la forme :

$$\mathbf{E}_\theta y(S) = 0 \quad \text{pour tout} \quad \theta \in \Theta \subset R^k.$$

THÉORÈME 2. *Une condition nécessaire et suffisante pour qu'une statistique S soit complète est que pour un $b_0(\theta)$ il existe un seul estimateur θ^* $\sigma(S)$-mesurable *) dans la classe de tous les estimateurs $\sigma(S)$-mesurables de K_{b_0}.*

S'il existe un seul estimateur $\sigma(S)$-mesurable dans K_{b_0}, il en sera de même dans toute autre classe K_b.

DÉMONSTRATION. Elle est évidente, puisque l'existence dans K_{b_0} de deux estimateurs $\sigma(S)$-mesurables $\theta_1^* = \varphi_1(S)$ et $\theta_2^* = \varphi_2(S)$ signifie que $\int \varphi_i(s)\mathbf{G}_\theta(ds) = b_0(\theta)$, $i = 1, 2$, et

$$\int [\varphi_1(s) - \varphi_2(s)]\, \mathbf{G}_\theta(ds) = 0 \quad \text{pour tout} \quad \theta \in \Theta,$$

de sorte que la complétude de S entraîne $\varphi_1(s) = \varphi_2(s)$ $[\mathscr{G}]$-presque partout. Réciproquement, supposons que $\int y(s)\mathbf{G}_\theta(ds) = 0$ pour tout $\theta \in \Theta$ et que $\theta_1^* = \varphi_1(s) \in K_b$. Alors $\theta_2^* = \varphi_1(s) + y(s) \in K_b$ et le fait qu'il existe un seul estimateur $\sigma(S)$-mesurable signifie que $y(s) = 0$ $[\mathscr{G}]$-presque partout. ◄

THÉORÈME 3. *Si une statistique exhaustive S est complète et $\theta^* \in K_b$, l'estimateur $\theta_S^* = \mathbf{E}_\theta(\theta^* | S)$ est le seul estimateur efficace de K_b.*

*) C'est-à-dire mesurable par rapport à la tribu $\sigma(S)$ engendrée par S et par suite pouvant se représenter par $\varphi(S)$, où φ est une fonction borélienne.

Ce théorème nous fournit des critères assez simples d'efficacité des estimateurs.

DÉMONSTRATION. D'après le théorème 2, il existe un seul estimateur $\sigma(S)$-mesurable dans K_b.

Soit θ^{**} un autre estimateur de K_b. Alors $\theta_S^{**} = \mathbf{E}_\theta(\theta^{**} | S) \in K_b$ et par suite $\theta_S^{**} = \theta_S^*$ $[\mathscr{P}]$-presque partout. De là et du théorème 1 il s'ensuit que

$$\mathbf{E}_\theta(\theta_S^* - \theta)^2 = \mathbf{E}_\theta(\theta_S^{**} - \theta)^2 \leqslant \mathbf{E}_\theta(\theta^{**} - \theta)^2,$$

et l'égalité n'est possible que pour $\theta^{**} = \theta_S^*$ p.s. ◄

COROLLAIRE 1. *Si S est une statistique exhaustive complète et θ^* un estimateur sans biais, alors θ_S^* est un estimateur efficace qui, de plus, est unique.*

EXEMPLE 2. Dans l'exemple 1 avec la distribution de Poisson, nous avons vu que pour $\lambda^* = x_1$

$$\lambda_S^* = \mathbf{E}_\lambda(x_1 | S) = \bar{x},$$

où $S = n\bar{x}$. Montrons que S est une statistique complète et, par suite, que $\bar{x}$ est un estimateur efficace. L'équation (2) pour la statistique S s'écrit

$$\sum_{k=0}^{\infty} y(k) e^{-n\lambda} \frac{(n\lambda)^k}{k!} = 0, \quad \forall \lambda \geqslant 0$$

ou, ce qui est équivalent,

$$v(z) = \sum y(k) \frac{z^k}{k!} = 0, \quad \forall z \geqslant 0. \tag{3}$$

Ce qui entraîne visiblement que $y(k) = 0$, puisque de la convergence de la série (3), disons pour $z = 1$, il s'ensuit que $v(z)$ est analytique pour $|z| < 1$ et identiquement nulle. Donc, les coefficients $y(k)$ de son développement en série sont nuls.

EXEMPLE 3. Soit $X \in \mathbf{U}_{0,\theta}$. Montrons que la statistique $S = x_{(n)} = \max_{i \leq n} x_i$ est complète. L'exhaustivité (et la minimalité) de S a été établie dans l'exemple 13.2. La distribution de S est définie par

$$\mathbf{P}(S < s) = (s/\theta)^n, \quad 0 \leqslant s \leqslant \theta,$$

de sorte que S admet une densité égale à $ns^{n-1}\theta^{-n}$ pour $s \in [0, \theta]$. L'équation (2) devient alors

$$\int_0^\theta y(s) \frac{ns^{n-1}}{\theta^n} ds = 0 \quad \text{pour} \quad \theta \in]0, \infty[.$$

De l'égalité $\int_0^\theta y(s)s^{n-1}ds = 0$ qui est valable pour tout θ, il s'ensuit visiblement que $y(s)s^{n-1} = 0$, $y(s) = 0$ presque partout.

Nous proposons au lecteur de vérifier si les statistiques exhaustives pour les autres familles paramétriques sont complètes. En particulier, établir que $\alpha^* = \frac{1}{\bar{x}}\left(1 - \frac{1}{n}\right)$ est le seul estimateur efficace du paramètre α de la famille $\Gamma_{\alpha,\,1}$ (cf. § 2).

Signalons maintenant que le théorème 3 nous suggère l'existence de relations entre les notions de complétude et de minimalité. A ce sujet, on a la proposition suivante qui, combinée aux théorèmes du § 13, nous fournit un critère de minimalité des statistiques exhaustives.

THÉORÈME 4. *Toute statistique exhaustive complète S est une statistique exhaustive minimale.*

DÉMONSTRATION. Soit $\mathfrak{A}_0$ une tribu exhaustive minimale (celle-ci existe en vertu du théorème 13.1). Supposons que $\mathbf{E}_\theta S$ existe et considérons la fonction $\psi = S - \mathbf{E}_\theta(S \mid \mathfrak{A}_0)$. Puisque $\mathfrak{A}_0 \subset \sigma(S)$, la fonction ψ sera $\sigma(S)$-mesurable et $\psi = \psi(S)$. Désignons par $\mathbf{G}_\theta$ la distribution de S. Pour tout θ on a alors de toute évidence $\mathbf{E}_\theta\psi(S) = 0$ ou ce qui est équivalent

$$\int \psi(s)\mathbf{G}_\theta(ds) = 0, \quad \forall\theta \in \Theta.$$

De là il s'ensuit en vertu de la complétude de S que $\psi(s) = 0$ $[\mathscr{G}]$-presque partout, $\mathscr{G} = \{\mathbf{G}_\theta\}$. Ceci exprime que $S = \mathbf{E}_\theta(S \mid \mathfrak{A}_0)$ $[\mathscr{P}]$-presque partout et par suite, S est mesurable par rapport *) à $\mathfrak{A}_0$, $\sigma(S) = \mathfrak{A}_0$.

Si $\mathbf{E}_\theta S$ n'existe pas, il faut à la place de S considérer la statistique Arctg S qui visiblement est équivalente à S par ses propriétés d'exhaustivité, de complétude et de minimalité. ◄

Signalons que la réciproque n'est pas vraie : *une statistique exhaustive minimale n'est pas nécessairement complète.* On pourrait citer des exemples correspondants dans le cas où la dimension l de la statistique est strictement supérieure à la dimension k du paramètre θ. Dans le § 13 nous avons vu par exemple que la densité conjointe de la statistique exhaustive minimale $S = (x_{(1)}, x_{(n)})$ pour la famille $\mathbf{U}_{\theta,\,1+\theta}$ est égale à

$$g_\theta(u, v) = \begin{cases} n(n-1)(v-u)^{n-2} & \text{si } u \geqslant \theta,\ v \leqslant 1+\theta,\ v > u, \\ 0 & \text{sinon.} \end{cases}$$

---

*) Par $\mathfrak{A}_0$ on comprendra ici la tribu complétée par les ensembles N tels que $\mathbf{P}_\theta(N) = 0$, $\forall\theta$.

Si l'on considère la fonction $y(u, v) = \varphi(v - u)$ et la transformation orthogonale $(v - u)/\sqrt{2} = t$, $(v + u)/\sqrt{2} = z$, l'intégrale (2) (étendue au triangle $u \geqslant \theta$, $v \leqslant 1 + \theta$, $v > u$) sera égale à

$$\int y(u, v) g_\theta(u, v) du\, dv = n(n - 1) \int_0^1 \varphi(x) x^{n-2}(1 - x) dx.$$

Il est évident que l'intégrale du second membre ne dépend pas de θ et il est aisé de choisir une fonction $\varphi(x) \not\equiv 0$ qui l'annulerait.

§ 15. Famille exponentielle

Supposons que $\theta = (\theta_1, \ldots, \theta_k)$ est un paramètre à k dimensions et que la densité $f_\theta(x)$ se représente sous la forme

$$f_\theta(x) = h(x) \exp\left\{ \sum_{j=1}^{k} a_j(\theta) U_j(x) + V(\theta) \right\}, \qquad (1)$$

où toutes les fonctions du second membre sont finies et mesurables.

DÉFINITION 1. On appellera *famille exponentielle* et on désignera par le symbole $\mathscr{E}$ toute famille de distributions $\{\mathbf{P}_\theta\}$ dont la densité est de la forme (1).

Pour rendre la représentation (1) la moins ambiguë possible, on admettra que *les fonctions* $a_0(\theta) \equiv 1$, $a_1(\theta), \ldots, a_k(\theta)$ *sont linéairement indépendantes sur* Θ.

Nous verrons que les familles exponentielles occupent une place privilégiée parmi les familles paramétriques de distributions, puisqu'elles permettent de nombreuses constructions générales de statistique mathématique sous forme explicite.

Les familles de distributions de forme plus particulière *) correspondant au cas où $a_j(\theta) = \theta_j$ sont parfois appelées familles exponentielles.

Comme exemples de familles exponentielles citons les familles de distributions $\{\mathbf{\Phi}_{\alpha, \sigma^2}\}$, $\{\mathbf{\Pi}_\lambda\}$, $\{\mathbf{B}_p\}$, $\{\mathbf{\Gamma}_{\alpha, \lambda}\}$, etc.

EXEMPLE 1. Considérons la distribution $\mathbf{\Gamma}_{\alpha, \lambda}$. Sa densité $\gamma_{\alpha, \lambda}(x)$ peut être mise sous la forme

$$\gamma_{\alpha, \lambda}(x) = \frac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda - 1} e^{-\alpha x} = x^{-1} \exp\left\{\lambda \ln x - \alpha x + \ln \frac{\alpha^\lambda}{\Gamma(\lambda)}\right\}, \quad x > 0,$$

---

*) Il s'agit en fait de la même chose : on est conduit à ce cas particulier en effectuant une application bijective $\gamma = \gamma(\theta)$, $\gamma = (\gamma_1, \ldots, \gamma_k)$ et en posant $\gamma_j = a_j(\theta)$.

de sorte qu'on peut poser ici

$$h(x) = \begin{cases} x^{-1}, & x > 0, \\ 0, & x \leqslant 0, \end{cases}$$

$$U_1(x) = \ln x, \quad U_2(x) = x, \quad V(\alpha, \lambda) = \ln \frac{\alpha^\lambda}{\Gamma(\lambda)},$$

$$a_1(\alpha, \lambda) = \lambda, \quad a_2(\alpha, \lambda) = -\alpha. \blacktriangleleft$$

La fonction de vraisemblance pour $X \in \mathbf{P} \in \mathscr{E}$ est égale à

$$f_\theta(X) = \exp\{(a(\theta), S) + nV(\theta)\} \prod_{i=1}^{n} h(x_i),$$

où

$$a(\theta) = (a_1(\theta), \ldots, a_k(\theta)), \; S = (S_1, \ldots, S_k),$$

$$S_j = S_j(X) = \sum_{i=1}^{n} U_j(x_i),$$

et (a, S) est le produit scalaire. De là et du théorème 12.1 il s'ensuit que S est une statistique exhaustive pour θ. On se propose de prouver que S est une statistique exhaustive minimale.

L'exponentielle de (1) est toujours strictement positive, puisque les fonctions $a_j(\theta)$, $U_j(x)$ et $V(\theta)$ sont finies. Ceci exprime que pour distribution $\mathbf{Q}$ dans le théorème 13.1 (distribution pour laquelle toutes les $\mathbf{P}_\theta$ sont absolument continues par rapport à $\mathbf{P_Q} = \int \mathbf{P}_t \mathbf{Q}(dt)$) on peut prendre une distribution concentrée en un point quelconque fixe θ^0. Le théorème 13.1 nous dit donc que la tribu $\mathfrak{A}_0$ engendrée par la fonction

$$r(X, \theta) = \frac{f_\theta(X)}{f_{\theta^0}(X)} = \exp\{(a(\theta) - a(\theta^0), S) + n(V(\theta) - V(\theta^0))\},$$

est une tribu exhaustive minimale.

THÉORÈME 1. *La statistique S est une statistique exhaustive minimale.*

DÉMONSTRATION. L'indépendance linéaire des fonctions $1, a_1(\theta), \ldots$ $\ldots a_k(\theta)$ sur Θ entraîne celle des fonctions $a_1(\theta) - a_1(\theta^0), \ldots, a_k(\theta) -$ $- a_k(\theta^0)$. Ceci exprime que dans Θ il existe k points $\theta^1, \ldots, \theta^k$ tels que les valeurs $a_{ij} = a_i(\theta^j) - a_i(\theta^0)$ forment une matrice A de déterminant non nul. Ceci exprime à son tour que les équations

$$(a(\theta^j) - a(\theta^0), S) = \ln r(X, \theta^j) - n(V(\theta^j) - V(\theta^0)), \; j = 1, \ldots, k,$$

admettent une seule solution S et par suite $\sigma(S) \subset \sigma(r(X, \theta_j)\,;\, j = 1, \ldots$ $\ldots, k) \subset \mathfrak{A}_0$. $\blacktriangleleft$

Dans l'exemple 1 nous avons étudié la distribution gamma et avons établi qu'elle était justiciable de la représentation (1) pour $\theta = (\alpha, \lambda)$ et

$$U_1(x) = \ln x, \qquad U_2(x) = x,$$

$$a_1(\alpha, \lambda) = \lambda, \qquad a_2(\alpha, \lambda) = -\alpha.$$

Il est évident que les conditions du théorème 1 sont remplies et la statistique $S = \left(\sum \ln x_i, \sum x_i\right)$ ou ce qui est équivalent la statistique $\left(\prod x_i, \sum x_i\right)$ est une statistique exhaustive minimale.

Si l'on renforce légèrement les conditions du théorème 1, la statistique S sera une statistique exhaustive complète, auquel cas sa minimalité peut être déduite de sa complétude.

THÉORÈME 2. *Soit $X \in \mathbf{P} \in \mathscr{E}$. Si une fonction a et un ensemble Θ sont tels que $a(\theta)$ balaye un parallélépipède à k dimensions lorsque θ parcourt Θ, alors S est une statistique exhaustive complète.*

Les conditions du théorème seront visiblement remplies pour le parallélépipède si l'ensemble Θ est « solide », c'est-à-dire contient des points intérieurs (ainsi que les sphères de R^k de rayon assez petit centrées en ces points), et les fonctions $a_j(\theta)$ sont linéairement indépendantes et différentiables au voisinage d'un point « solide » quelconque θ^0. Dans ces conditions, la transformation $a = a(\theta)$ envoie tout voisinage de θ^0 dans un ensemble solide.

Il est évident que l'exemple 1 de la distribution gamma vérifie les conditions du théorème 2, de sorte que la statistique $\left(\prod x_i, \sum x_i\right)$ est complète.

Le lecteur pourra vérifier aussi aisément que la statistique $\left(\sum x_i, \sum x_i^2\right)$ est une statistique exhaustive complète pour la distribution normale Φ_{α, σ^2}.

DÉMONSTRATION du théorème 2. Les fonctions $\psi(s, \theta)$ et $h(x)$ du théorème de factorisation de Neyman-Fisher sont ici

$$\psi(s, \theta) = \exp\{(a(\theta), s) + nV(\theta)\},$$

$$h(x) = \prod_{i=1}^{n} h(x_i).$$

Considérons sur $(R^k, \mathfrak{B}^k)$ la mesure indépendante de θ

$$\nu(B) = \int_{S^{-1}(B)} h(x)\mu^n(dx),$$

où $S^{-1}(B)$ est l'ensemble de tous les x tels que $S(x) \in B$.

Enonçons les deux propositions auxiliaires suivantes sous forme de lemmes.

LEMME 1. *La distribution* $\mathbf{G}_\theta(B) = \mathbf{P}_\theta(S(X)\in B)$ *de la statistique S est absolument continue par rapport à* ν *et admet au point s une densité égale à* $\psi(s, \theta)$.

DÉMONSTRATION. Elle découle de l'égalité

$$\mathbf{G}_\theta(B) = \int_{S(x)\in B} \psi(S(x), \theta)h(x)\mu^n(dx) = \int_{s\in B} \psi(s, \theta)\nu(ds),$$

qui est le résultat d'un changement de variables. ◄

LEMME 2. *Soient* $\mathbf{G}_1$ *et* $\mathbf{G}_2$ *deux mesures σ-finies dans* $(R^k, \mathfrak{B}^k)$. *Si* $\int e^{(a,u)}\mathbf{G}_1(du) = \int e^{(a,u)}\mathbf{G}_2(du)$ *existent pour tous les a d'un parallélépipède I de* R^k, *alors* $\mathbf{G}_1 = \mathbf{G}_2$.

DÉMONSTRATION. Pour simplifier les raisonnements, on se placera en dimension un ($k=1$) et on admettra que $I=\{x : |x| \leqslant \alpha\}$. Alors

$$h_j(a) = \int e^{au}\mathbf{G}_j(du), \quad j = 1, 2,$$

sont des fonctions analytiques pour $|a| < \alpha$. Par ailleurs, pour tout $b\in R$ sont définies les fonctions $h_j(z) = \int e^{(a+ib)u}\mathbf{G}_j(du)$ de la variable complexe $z=a+ib$. Il est évident que $h_j(z)$ seront analytiques dans la bande $|a| < \alpha$, $-\infty<b<\infty$. Puisque $h_1(z)=h_2(z)$ sur le segment de droite $b=0$, $|a| <\alpha$, il s'ensuit que $h_1(z)=h_2(z)$ pour tous les z de la bande mentionnée. Donc

$$\int e^{ibu}\mathbf{G}_1(du) = \int e^{ibu}\mathbf{G}_2(du). \tag{2}$$

A noter que les $\mathbf{G}_j$ peuvent être considérées comme des mesures de probabilité, puisque $h_j(0)=\int \mathbf{G}_j(du)<\infty$. Du théorème de correspondance biunivoque entre les fonctions caractéristiques et les distributions ([11]) et de (2), il s'ensuit que $\mathbf{G}_1=\mathbf{G}_2$.

Si le parallélépipède I est de la forme $\{x : |x-\alpha_0| \leqslant\alpha\}$, il convient de passer aux mesures $\mathbf{G}_j^*(du)=e^{\alpha_0 u}\mathbf{G}_j(du)$.

La démonstration est exactement la même en dimension $k>1$. ◄

Nous pouvons désormais passer directement à la *démonstration du théorème* 2.

Il nous faut prouver que si φ est une fonction mesurable dans $(R^k, \mathfrak{B}^k)$ et si existe

$$\int \varphi(s)\mathbf{G}_\theta(ds) = 0 \quad \text{pour tous les} \quad \theta\in\Theta, \tag{3}$$

alors $\varphi(s)=0$ $[\mathscr{S}]$-presque partout, $\mathscr{S}=\{\mathbf{G}_\theta\}_{\theta\in\Theta}$. Supposons que $\varphi=\varphi^+ - \varphi^-$, où $\varphi_\pm^+ \geqslant 0$. De (3) il résulte alors que $\int\varphi^+(s)\mathbf{G}_\theta(ds)=\int\varphi^-(s)\mathbf{G}_\theta(ds)$,

ou en vertu du lemme 1

$$\int \varphi^{+}(s)\psi(s,\theta)\nu(ds) = \int \varphi^{-}(s)\psi(s,\theta)\nu(ds),$$
$$\int \varphi^{+}(s)e^{(s,\,a(\theta))}\nu(ds) = \int \varphi^{-}(s)e^{(s,\,a(\theta))}\nu(ds).$$

Si $\nu^{\pm}(ds)=\varphi^{\pm}(s)\nu(ds)$, on obtient alors

$$\int e^{(s,\,a)}\nu^{+}(ds) = \int e^{(s,\,a)}\nu^{-}(ds)$$

pour tous les a d'un parallélépipède de R^k. Reste ensuite à appliquer le lemme 2. ◄

COROLLAIRE 1. *Si $X \in \mathbf{P} \in \mathscr{E}$, $\theta^* \in K_b$ et si sont remplies les conditions du théorème 2, alors l'estimateur $\theta_S^* = \mathbf{E}(\theta^* \mid S)$ est un estimateur efficace dans K_b.*

§ 16. Inégalité de Rao-Cramer et estimateurs R-efficaces

1. Inégalité de Rao-Cramer et ses conséquences. Dans les paragraphes précédents nous avons établi une série de critères d'efficacité des estimateurs. Mais ces critères revêtaient dans une certaine mesure un caractère qualitatif. Dans ce paragraphe nous poursuivons l'examen des estimateurs efficaces sous un point de vue légèrement différent. Voyons tout d'abord quelle est la plus petite valeur de l'erreur quadratique moyenne que l'on peut obtenir.

Etudions d'abord le cas où θ est un paramètre scalaire. Nous admettrons pour fixer les idées que l'ensemble Θ est un intervalle fini ou infini, fermé ou ouvert.

Pour répondre à la question posée, il nous faut imposer des conditions de régularité à $f_\theta(x)$. Soient comme précédemment

$$l(x,\theta) = \ln f_\theta(x), \quad L(X,\theta) = \sum_{i=1}^{n} l(x_i,\theta), \quad a(\theta) = \mathbf{E}_\theta\theta^* = \theta + b(\theta).$$

Supposons que sont remplies les conditions :

(R). *Les fonctions $\sqrt{f_\theta(x)}$ sont continûment dérivables par rapport à $\theta \in \Theta$ pour $[\mu]$-presque tous les x, et l'intégrale*

$$I(\theta) \equiv \int \frac{(f'_\theta(x))^2}{f_\theta(x)}\,\mu(dx) = \mathbf{E}_\theta[l'(x_1,\theta)]^2 \qquad (1)$$

existe, est strictement positive et continue par rapport à θ. (Ici et dans la suite le symbole prime désignera la dérivation par rapport à θ.)

Faisons la remarque suivante relativement à l'intégrale (1). Si x et son voisinage n'appartiennent pas au support $N_{\mathbf{P}_\theta} = \{x : f_\theta(x) > 0\}$ de la distribution $\mathbf{P}_\theta$, l'intégrant $(f'_\theta(x))^2/f_\theta(x)$ donne lieu à une indétermination de

type 0/0. On conviendra que ce rapport est nul. Nous adopterons la même convention pour la dérivée $l'(x, \theta) = f'_\theta(x)/f_\theta(x)$ lors de son intégration. On peut se passer de ces conventions si dès le départ on considère les intégrales de la forme $\mathbf{E}_\theta\varphi(\mathrm{x}_1, \theta)$ uniquement sur le domaine $N_{\mathbf{P}_\theta}$.

La fonction $I(\theta)$ s'appelle *quantité d'information de Fisher.* Elle joue un rôle très important en statistique mathématique et interviendra fréquemment dans la suite. Certaines de ses propriétés sont examinées au § 17.

Si l'ensemble Θ est compact, *la continuité de* $I(\theta)$ *dans les conditions* (*R*) *est équivalente à la condition.*

$$\sup_{\theta \in \Theta} \mathbf{E}_\theta([l'(\mathrm{x}_1, \theta)]^2 \; ; \; |l'(\mathrm{x}_1, \theta)| > N) \to 0$$

pour $N \to \infty$, que l'on pourrait appeler *convergence uniforme de l'intégrale* $I(\theta)$ (cf. Annexe VI).

On a l'inégalité suivante pour la variance des estimateurs θ^* de biais b.

THÉORÈME 1 (inégalité de Rao-Cramer). *Si* $\theta^* \in K_b$, *les conditions* (*R*) *sont satisfaites et* $\mathbf{E}_\theta(\theta^*)^2 < c < \infty$, alors

$$\mathbf{V}_\theta \theta^* \geqslant \frac{[1 + b'(\theta)]^2}{nI(\theta)} . \tag{2}$$

Si sur un intervalle $[\theta_1, \theta_2] \subset \Theta$ *l'égalité est réalisée dans* (2) *et* $\mathbf{V}_\theta\theta^* > 0$, *alors la fonction de vraisemblance* $f_\theta(X)$ *se représente pour* $\theta \in [\theta_1, \theta_2]$ *sous la forme*

$$f_\theta(X) = \exp\{\theta^* A(\theta) + B(\theta)\} h(X), \tag{3}$$

où $A(\theta)$ *et* $B(\theta)$ *ne dépendent pas de* X.

Réciproquement, si $\theta^* = \mathrm{const}$ *ou si l'on a la représentation* (3), *alors l'égalité est réalisée dans* (2).

La condition (3) exprime de toute évidence que la distribution de densité $f_\theta(x)$ dans $\mathscr{X}^n$ appartient à la famille exponentielle $\mathscr{E}$.

COROLLAIRE 1. *Si les conditions du théorème* 1 *sont satisfaites, alors*

$$\mathbf{E}_\theta(\theta^* - \theta)^2 \geqslant \frac{[1 + b'(\theta)]^2}{nI(\theta)} + b^2(\theta).$$

Pour tout estimateur sans biais θ^*, *on a*

$$\mathbf{E}_\theta(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)} .$$

Donc, dans les classes K_b, la plus petite valeur possible des erreurs quadratiques moyennes est définie par les seconds membres des inégalités écrites.

REMARQUE 1. Signalons à propos de la condition $\mathbf{E}_\theta(\theta^*)^2 < c < \infty$ que si $\mathbf{E}_\theta(\theta^*)^2 = \infty$, on a $\mathbf{V}_\theta\theta^* = \infty$ et l'inégalité (2) devient triviale. En vertu de (2), la condition $\mathbf{V}_\theta\theta^* > 0$ peut être remplacée par $(1 + b'(\theta))^2 > 0$.

REMARQUE 2. Outre les conditions (R), on peut indiquer d'autres conditions très voisines l'une de l'autre et assurant la réalisation du théorème 1. Nous nous attarderons sur celle d'entre elles qui nous sera utile dans les paragraphes suivants. Des conditions d'une forme différente seront exhibées dans le § 22.

Nous aurons besoin de la proposition auxiliaire suivante.

LEMME 1. *Supposons que les conditions* (R) *sont remplies et que* $S = S(X)$ *est une statistique quelconque telle que* $\mathbf{E}_\theta S^2 < c < \infty$ *pour* $\theta \in \Theta$. *Alors la fonction*

$$a_S(\theta) = \mathbf{E}_\theta S = \int S(x) f_\theta(x) \mu^n(dx) \qquad (4)$$

est dérivable par rapport à θ *et de plus*

$$a_S'(\theta) = \int S(x) f_\theta'(x) \mu^n(dx) = \mathbf{E}_\theta S L'(X, \theta). \qquad (5)$$

Cette proposition revêt un caractère technique et sa démonstration alourdirait considérablement l'exposé. Aussi l'ajournerons-nous à l'Annexe VI.

DÉMONSTRATION du théorème 1. En admettant que $S \equiv 1$ dans (5), on trouve que $a_S(\theta) \equiv 1$ et

$$\mathbf{E}_\theta L' = 0, \quad \mathbf{E}_\theta a(\theta) L' = 0. \qquad (6)$$

En utilisant encore (5) pour $S = \theta^*$ et (6), on obtient

$$\mathbf{E}_\theta \theta^* L' = a'(\theta), \quad \mathbf{E}_\theta(\theta^* - a(\theta)) L' = a'(\theta). \qquad (7)$$

L'inégalité de Cauchy-Bouniakovski nous donne

$$(a'(\theta))^2 \leqslant \mathbf{E}_\theta(\theta^* - a(\theta))^2 \mathbf{E}_\theta(L')^2 \qquad (8)$$

ou ce qui est équivalent

$$\mathbf{V}_\theta\theta^* \geqslant \frac{(1 + b'(\theta))^2}{\mathbf{E}_\theta(L')^2}. \qquad (9)$$

Vu que les variables aléatoires $l_j = l'(x_j, \theta)$ sont indépendantes, équidistribuées et admettent en vertu de (6) une espérance mathématique nulle, il vient $\mathbf{E}_\theta l_i l_j = 0$ pour $i \neq j$, $\mathbf{E}_\theta(L')^2 = \mathbf{E}_\theta\left(\sum_j l_j\right)^2 = \sum_{i,j} \mathbf{E}_\theta l_i l_j = n\mathbf{E}_\theta l_1^2 =$ $= nI(\theta)$. Ceci combiné à (9) prouve (2).

Prouvons la deuxième proposition du théorème. Pour simplifier, on admettra que Θ est confondu avec $[\theta_1, \theta_2]$ et que la mesure μ est concentrée sur la réunion des supports de $\mathbf{P}_\theta$, $\theta \in \Theta$. Dans (2) (ou dans (8)) l'égalité exprime que

$$\int (\theta^* - a(\theta)) f'_\theta(x) \mu^n(dx) =$$

$$= \left[\int (\theta^* - a(\theta))^2 f_\theta(x) \mu^n(dx) \int \frac{(f'_\theta(x))^2}{f_\theta(x)} \mu^n(dx) \right]^{1/2}$$

pour tous les $\theta \in \Theta$. La première intégrale du second membre étant par hypothèse strictement positive, cette égalité n'est possible que si

$$f'_\theta(x)/\sqrt{f_\theta(x)} = c(\theta)(\theta^* - a(\theta))\sqrt{f_\theta(x)} \quad [\mu^n]\text{-presque partout.} \tag{10}$$

Désignons par A l'ensemble des x pour lesquels est réalisée (10) et $|\theta^*| < \infty$. Alors $\mu(\bar{A}) = 0$ ($\bar{A}$ est le complémentaire de A). Fixons $x \in A$. La fonction $f_\theta(x)$ étant continue par rapport à θ, on a $f_t(x) > 0$ sur un intervalle $]t_1, t_2[\subset \Theta$ et en vertu de (10)

$$L'(x, \theta) = c(\theta)(\theta^* - a(\theta)) \tag{11}$$

sur cet intervalle. Remarquons maintenant que les relations (7), (11) et (2) entraînent

$$a'(\theta) = \mathbf{E}_\theta(\theta^* - a(\theta))L' = c(\theta)\mathbf{V}_\theta\theta^*, \quad \mathbf{V}_\theta\theta^* = \frac{(a'(\theta))^2}{nI(\theta)},$$

$$|c(\theta)| = \sqrt{\frac{nI(\theta)}{\mathbf{V}_\theta\theta^*}}, \tag{12}$$

de sorte que $\mathbf{V}_\theta\theta^*$ est continue par rapport à θ avec $a'(\theta)$ et $I(\theta)$, quant à $|c(\theta)|$ elle est uniformément bornée avec $a(\theta)$ sur $[\theta_1, \theta_2]$; il en est de même de la dérivée $L'(x, \theta)$ dans (11). Or ceci exprime que $L(x, t)$ est finie, $f_\theta(x) > 0$ partout sur $\Theta = [\theta_1, \theta_2]$, de sorte que (11) est satisfaite pour tous les θ. En intégrant (11) entre θ_1 et θ_2, on obtient

$$L(x, \theta) = \theta^* \int_{\theta_1}^{\theta} c(t)dt - \int_{\theta_1}^{\theta} c(t)a(t)dt + L(x, \theta_1),$$

ce qui est équivalent à (3) pour $[\mu^n]$-presque tous les x. Ceci prouve (3), puisque le changement de $f_\theta(x)$ sur un ensemble μ^n-négligeable est sans effet.

Considérons maintenant la dernière assertion du théorème. Si $\theta^* = \text{const}$, alors $b'(\theta) = -1$ et les deux membres de l'inégalité (2) sont nuls. Supposons maintenant que (3) est remplie. En dérivant la fonction $L(X, \theta)$ par rapport à θ, on obtient alors

$$L'(X, \theta) = \theta^* A'(\theta) + B'(\theta).$$

De (7) il s'ensuit que $a(\theta)A'(\theta) + B'(\theta) = 0$. Donc,

$$L'(X, \theta) = A'(\theta)(\theta^* - a(\theta))$$

et par suite (cf. (10)), l'égalité est réalisée dans (2). ◄

Dans la suite nous omettrons le cas trivial $\theta^* =$ const et supposerons que $\mathbf{V}_\theta\theta^* > 0$ partout sur Θ. On a alors le

COROLLAIRE 2. *Les conditions* (R) *étant remplies, pour que la borne inférieure soit atteinte dans l'inégalité de Rao-Cramer, il est nécessaire et suffisant que l'estimateur* θ^* *soit exhaustif et que la fonction* $\psi(\theta^*, \theta)$ *de l'égalité de factorisation soit de la forme*

$$\psi(\theta^*, \theta) = \exp\{\theta^* A(\theta) + B(\theta)\},$$

où $A(\theta)$ *et* $B(\theta)$ *sont des fonctions dérivables.*

COROLLAIRE 3. *Si les conditions* (R) *sont remplies,* $\theta^* \in K_b$ *et l'égalité est réalisée dans l'inégalité de Rao-Cramer, alors* θ^* *est un estimateur efficace dans* K_b.

Cette proposition résulte de la représentation

$$\mathbf{E}_\theta(\theta^* - \theta)^2 = \mathbf{V}_\theta\theta^* + b^2(\theta).$$

A noter que la réciproque est généralement mise en défaut : un estimateur peut être efficace dans K_b sans que la borne inférieure $\dfrac{(1 + b'(\theta))^2}{nI(\theta)}$ de la variance soit atteinte.

EXEMPLE 1. Soit $X \in \Gamma_{\alpha,1}$. Ici $f_\alpha(X) = \alpha^n e^{-\alpha n\bar{x}}$. Les conditions ($R$) sont remplies dans le domaine $\Theta \subseteq \{\alpha \geqslant \delta > 0\}$. Il est évident que $S = n\bar{x}$ est une statistique exhaustive complète. Donc, l'estimateur $\alpha^* = \bar{x}^{-1} = \mathbf{E}_\alpha(\bar{x}^{-1} \mid S)$ est un estimateur efficace dans la classe K_b de biais $b(\alpha) = \mathbf{E}_\alpha\bar{x}^{-1} - \alpha$.

Remarquons maintenant que $S \in \Gamma_{\alpha,n}$ de sorte que pour $n > 1$ (cf. § 2), on a $\mathbf{E}_\alpha\bar{x}^{-1} = n\mathbf{E}_\alpha S^{-1} = \dfrac{n}{n-1}\alpha$.

L'estimateur $\alpha^{**} = \dfrac{n-1}{n\bar{x}} = \alpha^*\left(1 - \dfrac{1}{n}\right)$ sera donc sans biais pour $n > 1$.

De façon analogue, pour $n > 2$ on trouve (cf. § 2, ainsi que l'exemple 4.1)

$$\mathbf{E}_\alpha(\alpha^{**})^2 = (n-1)^2\mathbf{E}_\alpha S^{-2} = \frac{n-1}{n-2}\alpha^2,$$

$$\mathbf{V}_\alpha\alpha^{**} = \alpha^2\left[\frac{n-1}{n-2} - 1\right] = \frac{\alpha^2}{n-2}.$$

Ainsi, l'estimateur α^{**} est efficace pour $n>2$. Cependant, le critère (3) n'est pas rempli, puisque

$$f_\alpha(X) = \alpha^{-n}e^{-\alpha(n-1)/\alpha^{**}}.$$

Donc, la borne inférieure n'est pas atteinte dans l'inégalité de Rao-Cramer. On peut s'en assurer directement. En effet, on a ici $l(x, \alpha) = \ln\alpha - \alpha x$; $l'(x, \alpha) = 1/\alpha - x$ et

$$I(\alpha) = \mathsf{E}_\alpha[l'(x_1, \alpha)]^2 = \mathsf{E}_\alpha\left(\frac{1}{\alpha} - x_1\right)^2 = \frac{1}{\alpha^2} - \frac{2}{\alpha^2} + \frac{2}{\alpha^2} = \frac{1}{\alpha^2}.$$

Donc, pour $n>2$
$$\frac{1}{nI(\theta)} = \frac{\alpha^2}{n} < \frac{\alpha^2}{n-2} = \mathsf{V}_\alpha\alpha^{**}.$$

Par conséquent, la réalisation de la borne inférieure dans (2) est une condition plus astreignante que l'efficacité.

2. Estimateurs *R*-efficaces et asymptotiquement *R*-efficaces. Supposons remplies les conditions (R). Dans ce cas, la réalisation (exacte ou asymptotique) de la borne inférieure dans l'inégalité de Rao-Cramer peut servir d'important critère de qualité des estimateurs, un critère qui est étroitement lié à la notion d'efficacité.

DÉFINITION 1. On appelle *estimateur R-efficace* (ou *régulièrement efficace*) *dans la classe* K_b un estimateur θ^* tel que

$$\mathsf{E}_\theta(\theta^* - \theta)^2 = \frac{(1 + b'(\theta))^2}{nI(\theta)} + b^2(\theta).$$

On appellera tout simplement *R-efficace* un estimateur R-efficace dans la classe K_0 des estimateurs sans biais.

On dit qu'un estimateur θ^* est *asymptotiquement R-efficace* si

$$\mathsf{E}_\theta(\theta^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

On remarque que, contrairement aux définitions du § 8 qui revêtaient un caractère plus qualitatif, les définitions de la R-efficacité reposent sur la comparaison avec des valeurs numériques connues reliées essentiellement à la quantité d'information de Fisher ou plus exactement à la quantité $(nI(\theta))^{-1}$.

Pour qu'un estimateur θ^* soit R-efficace, il est nécessaire et suffisant que soit réalisée (3).

De ce qui précède il s'ensuit que les estimateurs R-efficaces sont efficaces, mais la réciproque n'est pas vraie ; les estimateurs R-efficaces sont tout simplement plus rares, ce qui est un défaut de la borne inférieure dans l'inégalité de Rao-Cramer, mais pas des estimateurs.

En statistique mathématique, les estimateurs R-efficaces sont appelés tout simplement efficaces. Pour notre part, il nous semble plus naturel de réserver le terme « efficace » à des estimateurs meilleurs dans une acception plus large (cf. définition 8.1).

THÉORÈME 2. *Si les conditions* (R) *sont remplies et si existe un estimateur R-efficace, alors il est confondu avec un estimateur du maximum de vraisemblance.*

DÉMONSTRATION. Nous savons que la réalisation de (3) entraîne l'égalité (cf. (11))

$$L'(X, \theta) = (\theta^* - \theta)c(\theta).$$

Par ailleurs, puisque $b(\theta) = 0$, il vient de (12)

$$c(\theta) = 1/\mathbf{V}_\theta\theta^* = nI(\theta) > 0,$$

quel que soit $\theta \in \Theta$. Ceci exprime que $L'(X, \theta) < 0$ pour $\theta > \theta^*$ et $L'(X, \theta) > 0$ pour $\theta < \theta^*$. Donc, $L(X, \theta)$ atteint son maximum pour $\theta = \theta^*$. ◄

L'exemple 1 ci-dessus montre que contrairement aux estimateurs R-efficaces, les estimateurs efficaces peuvent ne pas être confondus avec ceux du maximum de vraisemblance. Dans cet exemple, $\bar{x}^{-1}$ est un estimateur du maximum de vraisemblance, alors que $\frac{n-1}{n}(\bar{x})^{-1}$ est un estimateur efficace. Ces deux estimateurs sont visiblement asymptotiquement R-efficaces.

Considérons la classe $\tilde{K}_0$ des estimateurs θ^* tels que pour $n \to \infty$ et tout $\theta \in \Theta$

$$|b(\theta)| \leqslant \epsilon(\theta, n)/\sqrt{n}, \quad |b'(\theta)| \leqslant \epsilon(\theta, n),$$

$$\mathbf{E}_0(\theta^*)^2 < c < \infty,$$

où $\epsilon(\theta, n)$ est une fonction telle que $\epsilon(\theta, n) = o(1)$ pour $n \to \infty$.

Chaque classe $\tilde{K}_0$ est remarquable par le fait que dans l'inégalité de Rao-Cramer, la borne inférieure est de la forme $(1 + o(1))/[nI(\theta)]$. Au § 20 nous verrons que dans bien des cas, en cherchant des estimateurs asymptotiquement optimaux, on peut se borner à étudier des estimateurs θ^* appartenant à de telles classes.

THÉORÈME 3. *Si les conditions* (R) *sont réunies, tout estimateur asymptotiquement R-efficace de* $\tilde{K}_0$ *est un estimateur asymptotiquement efficace dans* $\tilde{K}_0$.

DÉMONSTRATION. Elle est évidente : si θ_1^* est un estimateur asymptotiquement R-efficace, alors

$$\mathbf{E}_\theta(\theta_1^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

Par ailleurs, comme déjà signalé, en vertu de l'inégalité de Rao-Cramer, pour tous les $\theta^* \in \tilde{K}_0$

$$\liminf_{n\to\infty} \mathbf{E}_\theta n(\theta^* - \theta)^2 \geqslant I^{-1}(\theta) = \lim_{n\to\infty} \mathbf{E}_\theta n(\theta_1^* - \theta)^2. \blacktriangleleft$$

Il est également clair que si existe un estimateur asymptotiquement R-efficace, tout estimateur asymptotiquement efficace dans $\tilde{K}_0$ sera asymptotiquement R-efficace.

Nous verrons plus bas (cf. § 25) que sous certaines hypothèses complémentaires, les estimateurs asymptotiquement R-efficaces existent toujours et par suite, le théorème 3 admet une réciproque, savoir que tout estimateur asymptotiquement efficace de $\tilde{K}_0$ est asymptotiquement R-efficace, c'est-à-dire que $\mathbf{E}_\theta(\theta^* - \theta)^2 \sim [nI(\theta)]^{-1}$.

THÉORÈME 4. *Supposons remplies les conditions* (R). *Si* θ_1^* *et* θ_2^* *sont des estimateurs asymptotiquement R-efficace de* $\tilde{K}_0$, *ils sont asymptotiquement équivalents au sens suivant :*

$$\sqrt{n}(\theta_1^* - \theta_2^*) \underset{P}{\to} 0.$$

DÉMONSTRATION. Elle est calquée sur celle du théorème 8.2. Puisque $\theta^* = (\theta_1^* + \theta_2^*)/2 \in \tilde{K}_0$, en vertu de (8.11) et de l'inégalité de Rao-Cramer, il vient

$$\limsup_{n\to\infty} \mathbf{E}_\theta n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \blacktriangleleft$$

EXEMPLE 2. L'estimateur $\alpha^* = \bar{x}$ de la moyenne α de la distribution normale Φ_{α,σ^2}, σ^2 étant connue, est un estimateur R-efficace. On s'en assure sans peine en vérifiant par exemple la condition (3). Un autre moyen consiste à comparer $\mathbf{V}_\alpha \alpha^* = \sigma^2/n$ à la plus petite valeur possible $(nI(\alpha))^{-1}$ des variances des estimateurs sans biais. Il vient

$$l(x, \alpha) = -\ln\sqrt{2\pi}\,\sigma - (x - \alpha)^2/(2\sigma^2),$$
$$l'(x, \alpha) = (x - \alpha)/\sigma^2,$$
$$I(\alpha) = \mathbf{E}_\alpha[l'(x_1, \alpha)]^2 = \mathbf{E}_\alpha(x_1 - \alpha)^2/\sigma^4 = 1/\sigma^2,$$

de sorte que $\mathbf{V}_\alpha \alpha^* = (nI(\alpha))^{-1} = \sigma^2/n$.

EXEMPLE 3. Considérons l'estimateur $\theta^* = S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2$ du paramètre $\theta = \sigma^2$ de la distribution normale, α étant connu. On trouve sans

peine que $\mathbf{V}_\theta\theta^* = \mathbf{E}_\theta(\theta^* - \sigma^2)^2 = 2\sigma^4/n$. Par ailleurs,

$$l'(x_1, \theta) = -\frac{1}{2\theta} + \frac{(x_1 - \alpha)^2}{2\theta^2},$$

$$I(\theta) = \mathbf{E}_\theta[l'(x_1, \theta)]^2 = \frac{1}{4\theta^4}\mathbf{E}_\theta[(x_1 - \alpha)^2 - \theta]^2 = \frac{1}{2\theta^2} = \frac{1}{2\sigma^4}.$$

Donc, ici aussi $\mathbf{V}_\theta\theta^* = (nI(\theta))^{-1}$ et l'estimateur $\theta^* = S_1^2$ est R-efficace.

La variance de l'estimateur sans biais $S_0^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ est égale à $\frac{2\sigma^4}{n-1}$, de sorte que cet estimateur n'est ni R-efficace ni tout simplement efficace. Par ailleurs, il est évident que S_0^2 est asymptotiquement R-efficace.

Si au lieu d'estimer σ^2 on estime le paramètre $\theta = \sigma$, on n'obtiendra pas d'estimateur R-efficace. Un estimateur sans biais de σ sera

$$\sigma^* = \sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}S,$$

puisque

$$\mathbf{E}_\sigma S = \frac{\sigma}{\sqrt{n}}\mathbf{E}_\sigma\sqrt{\frac{1}{\sigma^2}\sum(x_i - \alpha)^2}$$

où $\frac{nS^2}{\sigma^2} = \frac{1}{\sigma^2}\sum(x_i - \alpha)^2$ suit la distribution $\mathbf{H}_n = \Gamma_{1/2,\, n/2}$, donc (cf. § 2)

$$\mathbf{E}_\sigma S = \frac{\sigma\sqrt{2}}{\sqrt{n}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}, \quad \mathbf{E}_\sigma\sigma^* = \sigma.$$

La statistique S étant minimale et exhaustive complète, l'estimateur σ^* est efficace. La formule de Stirling nous permet de nous assurer sans peine que $\sigma^* = S(1 + O(1/n))$.

Comparons maintenant $\mathbf{V}_\sigma\sigma^*$ à la borne inférieure $(nI(\sigma))^{-1}$. On a

$$\mathbf{V}_\sigma\sigma^* = \mathbf{E}_\sigma\left(\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}S - \sigma\right)^2 = \sigma^2\left[\frac{n}{2}\frac{\Gamma^2\left(\frac{n}{2}\right)}{\Gamma^2\left(\frac{n+1}{2}\right)} - 1\right]. \quad (13)$$

Par ailleurs,

$$l'(x, \sigma) = -\frac{1}{\sigma} + \frac{(x - \alpha)^2}{\sigma^3},$$

$$I(\sigma) = \mathbf{E}_\sigma[l'(x_1, \sigma)]^2 = \frac{1}{\sigma^6}\mathbf{E}_\sigma[(x_1 - \alpha)^2 - \sigma^2]^2 = \frac{2}{\sigma^2},$$

de sorte que $(nI(\sigma))^{-1} = \sigma^2/(2n)$. Or cette valeur est différente de (13). Pour $n=3$ par exemple, leur rapport est égal à 0,936. Donc, il n'existe pas d'estimateurs R-efficaces ici. Lorsque $n \to \infty$, le coefficient de σ^2 dans (13) se conduit comme $\frac{1}{2n} + O\left(\frac{1}{n^2}\right)$, de sorte que σ^* est un estimateur asymptotiquement R-efficace.

3. Inégalité de Rao-Cramer dans le cas vectoriel. Dans ce numéro $\theta = (\theta_1, \ldots, \theta_k)$ et l'estimateur $\theta^* = (\theta_1^*, \ldots, \theta_k^*)$ sont des vecteurs à k dimensions. Posons comme précédemment

$$a(\theta) = \mathbf{E}_\theta \theta^* = \theta + b(\theta), \quad b(\theta) = (b_1(\theta), \ldots, b_k(\theta))$$

et considérons les classes K_b d'estimateurs de biais $b(\theta)$ fixé.

Les conditions (R) se généralisent de la manière suivante au cas vectoriel. Posons

$$l(x, \theta) = \log f_\theta(x), \quad l_i'(x, \theta) = \frac{\partial}{\partial \theta_i} l(x, \theta),$$

$$I_{ij}(\theta) = \mathbf{E}_\theta l_i'(\mathrm{x}_1, \theta) l_j'(\mathrm{x}_1, \theta)$$

et supposons que sont remplies les conditions

(R). *Les fonctions* $\sqrt{f_\theta(x)}$ *sont continûment dérivables par rapport à* θ_j pour $[\mu]$-*presque toutes les valeurs de* x. *La matrice*

$$I(\theta) = \| I_{ij}(\theta) \|,$$

$$I_{ij}(\theta) = \int l_i'(x\,;\theta) l_j'(x\,;\theta) f_\theta(x) \mu(dx)$$

est continue par rapport à θ *) *et son déterminant* $|I(\theta)|$ *est non nul.*

Vu que $I(\theta)$ est la matrice des moments d'ordre deux $\mathbf{E}_\theta l_i l_j$ des variables aléatoires $l_i = l_i'(\mathrm{x}_1, \theta)$, elle est définie positive, puisque pour tout vecteur $\alpha = (\alpha_1, \ldots, \alpha_k) \neq 0$ on a

$$\sum \alpha_i \alpha_j \mathbf{E}_\theta l_i l_j = \mathbf{E}_\theta \left(\sum \alpha_i l_i\right)^2 \geqslant 0,$$

où l'égalité à zéro est exclue par la condition $|I(\theta)| \neq 0$.

Comme précédemment, l'inégalité matricielle $\sigma_1^2 \geqslant \sigma_2^2$ sera comprise comme l'inégalité $\alpha \sigma_1^2 \alpha^T \geqslant \alpha \sigma_2^2 \alpha^T$ pour tout vecteur ligne $\alpha = (\alpha_1, \ldots, \alpha_k) \neq 0$. Ceci équivaut de toute évidence à la semi-définition positive de la matrice $\sigma_1^2 - \sigma_2^2$. L'inégalité stricte correspondra à la définition positive, de sorte que par exemple $I(\theta) > 0$.

---

*) Il suffit d'exiger que $I_{ii}(\theta)$ soit uniformément convergente (cf. Annexe VI).

THÉORÈME 1A. *Si $\theta^* \in K_b$ et si sont remplies les conditions (R), la matrice des moments d'ordre deux $\sigma^2 = \|\sigma_{ij}\| = \mathbf{E}_\theta(\theta^* - a(\theta))^T(\theta^* - a(\theta))$ de tout estimateur θ^* du vecteur ligne θ vérifie l'inégalité*

$$\sigma^2 \geqslant \frac{1}{n}(E + D(\theta))I^{-1}(\theta)(E + D(\theta))^T \tag{14}$$

où E est la matrice unité, $D(\theta) = \|b_{ij}(\theta)\|$, $b_{ij}(\theta) = \dfrac{\partial b_i(\theta)}{\partial \theta_j}$.

Supposons que $|\sigma^2| > 0$ (ou $|E + D(\theta)| > 0$) pour tous les θ. Dans ce cas, l'égalité est réalisée dans (14) *si et seulement si la distribution de l'échantillon appartient à une famille exponentielle de type spécial, c'est-à-dire lorsque pour des fonctions scalaires $B(\theta)$ et $h(X)$, on a*

$$f_\theta(X) = \exp\{(\theta^*, A(\theta)) + B(\theta)\}h(X), \tag{15}$$

où le vecteur $A(\theta) = (A_1(\theta), \ldots, A_k(\theta))$ admet une matrice de dérivées égale à

$$\|A_{ij}\| \equiv \left\| \frac{\partial A_i(\theta)}{\partial \theta_j} \right\| = n[(E + D(\theta))^{-1}]^T I(\theta).$$

Pour les estimateurs θ^* sans biais, il est évident que

$$\sigma^2 \geqslant (nI(\theta))^{-1}$$

et l'égalité n'est possible que si a lieu (15), *où $\|A_{ij}\| = nI(\theta)$.*

Si donc l'on réussit à trouver un estimateur θ^* sans biais de matrice des moments d'ordre deux $[nI(\theta)]^{-1}$, alors cet estimateur sera efficace.

Signalons que

$$\mathbf{E}_\theta(\theta^* - \theta)^T(\theta^* - \theta) = \sigma^2 + b^T(\theta)b(\theta).$$

DÉMONSTRATION du théorème 1A. Posons

$$L_j' = L_j'(X, \theta) = \sum_{i=1}^{n} l_j'(x_i, \theta),$$

$$L' = L'(X, \theta) = (L_1', \ldots, L_k').$$

Exactement comme pour le cas scalaire, on trouve que

$$\mathbf{E}_\theta l_j'(x_1, \theta) = 0, \quad \mathbf{E}_\theta \theta_i^* L_j'(X, \theta) = 1 + b_{ij}(\theta),$$

où $b_{ij}(\theta)$ sont continues, ou, ce qui revient au même, que

$$\mathbf{E}_\theta L' = 0, \tag{16}$$

$$\mathbf{E}_\theta(\theta^*)^T L' = E + D(\theta), \tag{17}$$

où la matrice $D(\theta)$ est continue. De là il s'ensuit

$$\mathsf{E}_\theta(\theta^* - a(\theta))^T L' = E + D(\theta). \tag{18}$$

Prouvons maintenant l'inégalité suivante (la version matricielle de l'inégalité de Cauchy-Bouniakovski).

LEMME 2. *Soient ξ et η des matrices de même dimension (pas nécessairement carrées) à éléments aléatoires. Si la matrice $\mathsf{E}\eta\eta^T$ est inversible, alors*

$$\mathsf{E}\xi\xi^T \geqslant \mathsf{E}\xi\eta^T(\mathsf{E}\eta\eta^T)^{-1}\mathsf{E}\eta\xi^T. \tag{19}$$

Ceci étant, l'égalité n'est possible que si $\xi = z\eta$, $z = \mathsf{E}\xi\eta^T(\mathsf{E}\eta\eta^T)^{-1}$.

DÉMONSTRATION. Puisque toute matrice A vérifie l'inégalité $AA^T \geqslant 0$ (AA^T est semi-définie positive), il vient

$$0 \leqslant \mathsf{E}(\xi - z\eta)(\xi - z\eta)^T = \mathsf{E}\xi\xi^T - z\mathsf{E}\eta\xi^T - \mathsf{E}\xi\eta^T z^T + z\mathsf{E}\eta\eta^T z^T.$$

En posant $z = \mathsf{E}\xi\eta^T(\mathsf{E}\eta\eta^T)^{-1}$, on obtient l'inégalité annoncée.

La proposition concernant les conditions d'égalité dans (19) est évidente. ◄

Revenons à la *démonstration du théorème* 1A. Posons $\xi = (\theta^* - a(\theta))^T$, $\eta = (L')^T$ dans (19). Alors

$$\mathsf{E}_\theta\xi\xi^T = \mathsf{E}_\theta(\theta^* - a(\theta))^T(\theta^* - a(\theta)) = \sigma^2.$$

De (16) et de l'indépendance des x_i on déduit que

$$\mathsf{E}_\theta\eta\eta^T = \mathsf{E}_\theta(L')^T L' = nI(\theta).$$

De (18) il vient enfin

$$\mathsf{E}_\theta\xi\eta^T = \mathsf{E}_\theta(\theta^* - a(\theta))^T L' = E + D(\theta).$$

Ce qui prouve l'inégalité (14).

Dans (14) l'égalité n'est possible, en vertu du lemme 2, que si pour les points (x, θ) tels que $f_\theta(x) > 0$, l'on a

$$(\theta^* - a(\theta))^T = (E + D(\theta))(nI(\theta)^{-1}(L')^T$$

ou, ce qui est équivalent,

$$L' = (\theta^* - a(\theta))n[(E + D(\theta))^{-1}]^T I(\theta). \tag{20}$$

Remarquons maintenant que l'égalité dans (14) entraîne

$$|E + D(\theta)|^2 = n|\sigma^2| \cdot |I(\theta)|,$$

et la non-nullité du déterminant $|\sigma^2|$ entraîne celle de $|E + D(\theta)|$, ce qui exprime qu'existe la matrice inverse uniformément bornée $(E + D(\theta))^{-1}$.

Donc, la dérivée L' de (20) sera bornée, $f_\theta(x)>0$ partout sur Θ et l'égalité (20) sera vérifiée partout sur Θ. Si maintenant s est un chemin quelconque reliant les points θ_1 et θ_2 du domaine Θ, alors

$$L(X,\ \theta) = \int_s (L',\ ds\ ) + L(X,\ \theta_0),$$

où ds est un élément vectoriel de chemin ; $(L', ds)=(L', s'(l))dl$ l'accroissement de $L(X, \theta)$ sur le chemin s ; l la « longueur » du chemin parcouru. Donc, en vertu de (20),

$$L(X,\ \theta) = \theta^* A(\theta) + B(\theta) + H(X), \tag{21}$$

où $B(\theta)$ et $H(X)$ sont des fonctions scalaires, $A(\theta)=(A_1(\theta), \ldots, A_k(\theta))$ un vecteur dépendant seulement de ses arguments. Ceci prouve (15).

Si (21) est réalisée, on a

$$L' = \theta^* \| A_{ij} \| + B'(\theta),$$

où, en vertu de l'égalité $\mathbf{E}_\theta L'=0$,

$$B'(\theta) = - a(\theta)\| A_{ij} \|.$$

En multipliant les deux membres de l'égalité $L'=(\theta^*-a(\theta))\|A_{ij}\|$ à gauche par $(\theta^*-a(\theta))^T$, on trouve en vertu de (18) que, pour que la condition (20) qui exprime l'égalité dans (14) soit réalisée, il faut que

$$\| A_{ij} \| = n[(E + D(\theta))^{-1}]^T I(\theta). \blacktriangleleft$$

Toutes les remarques relatives à l'inégalité de Rao-Cramer ainsi que la définition de la R-efficacité dans le cas scalaire sont valables dans le cas vectoriel *mutatis mutandis.*

En particulier, on appellera *estimateurs asymptotiquement R-efficaces* les estimateurs θ^* tels que

$$\mathbf{E}_\theta(\theta^* - \theta)^T(\theta^* - \theta) = \sigma^2 + b^T(\theta)b(\theta) = (nI(\theta))^{-1} + o(1/n).$$

L'analogue du théorème 2 s'énonce comme suit.

THÉORÈME 2A. *Supposons remplies les conditions* (R). *Si* θ^* *est un estimateur R-efficace, il est estimateur du maximum de vraisemblance.*

DÉMONSTRATION. Pour prouver qu'un estimateur R-efficace est le seul point de maximum, il suffit de s'assurer que $L'(X, \theta^*)=0$ et que

$$(\text{grad } L(X,\ \theta),\ u) = (L'(X,\ \theta),\ \theta - \theta^*) < 0$$

pour $\theta=\theta^*+u$, $u\neq 0$. Or s'il existe un estimateur R-efficace, on a (cf. (20))

$$L'(X,\ \theta) = (\theta^* - \theta)nI(\theta),$$

d'où l'on déduit immédiatement les deux relations annoncées. La deuxième de ces relations résulte du fait que

$$(L', u) = -unI(\theta)u^T,$$

où $uI(\theta)u^T$ est une forme quadratique définie positive. ◄

EXEMPLE 4. Considérons une famille à deux paramètres de distributions normales Φ_{α,σ^2}. Cette famille est exponentielle, puisque (ici $\theta = (\theta_1, \theta_2)$, $\theta_1 = \alpha$, $\theta_2 = \sigma^2$)

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} - \ln\sigma\right\}.$$

L'estimateur $\theta^* = (\theta_1^*, \theta_2^*)$, où $\theta_1^* = \bar{x}$, $\theta_2^* = S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum x_i^2 - \bar{x}^2\right)$, est efficace, puisqu'il appartient à K_0 et la statistique $\left(\sum x_i, \sum x_i^2\right)$ est, comme nous l'avons vu au § 15, une statistique exhaustive complète (cf. théorème 14.4).

L' estimateur du maximum de vraisemblance $\left(\bar{x}, \frac{1}{n}\sum(x_i - \bar{x})^2\right)$ diffère de θ^* seulement par un facteur multiplicatif $\frac{n-1}{n}$ en la deuxième coordonnée, ce qui en fait un estimateur à biais. La fonction $f_\theta(X)$ n'admettra pas la représentation exponentielle spéciale (15) pour l'estimateur choisi θ^*, puisque

$$f_\theta(X) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum x_i^2 + \frac{\alpha}{\sigma^2}\sum x_i - \frac{n\alpha^2}{2\sigma^2} n \ln\sigma\right\} =$$

$$= (2\pi)^{-n/2} \exp\left\{\frac{\alpha n}{\sigma^2}\theta_1^* - \frac{n-1}{2\sigma^2}\theta_2^* - \frac{n}{2\sigma^2}(\theta_1^*)^2 - \frac{n\alpha^2}{2\sigma^2} - n\ln\sigma\right\}.$$

Ce qui exprime que la borne inférieure ne sera pas atteinte dans l'inégalité de Rao-Cramer pour le cas vectoriel.

Le plus petit ellipsoïde de dispersion défini en vertu du théorème 1A par la matrice $I(\theta)$ (ou $I^{-1}(\theta)$) ne sera atteint qu'asymptotiquement pour $n \to \infty$, de sorte que l'estimateur θ^*, à défaut d'être R-efficace, sera asymptotiquement R-efficace. Vérifions-le directement.

Calculons d'abord la matrice $I(\theta)$. On a

$$l_1'(x, \theta) = \frac{(x-\alpha)}{\sigma^2}, \quad l_2'(x, \theta) = \frac{(x-\alpha)^2}{2\sigma^4} - \frac{1}{2\sigma^2}$$

(on rappelle que l_2' est la dérivée par rapport à σ^2 et non pas à σ ; comparer avec l'exemple 3). Donc

$$I_{11}(\theta) = \mathbf{E}_\theta \frac{(x_1 - \alpha)^2}{\sigma^4} = \frac{1}{\sigma^2},$$

$$I_{12}(\theta) = I_{21}(\theta) = \mathbf{E}_\theta\left[\frac{(x_1 - \alpha)^3}{2\sigma^6} - \frac{x_1 - \alpha}{2\sigma^4}\right] = 0,$$

$$I_{22}(\theta) = \frac{1}{4\sigma^8} \mathbf{E}_\theta[(x_1 - \alpha)^2 - \sigma^2]^2 = \frac{1}{2\sigma^4}.$$

D'où il vient

$$(nI(\theta))^{-1} = \begin{vmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{vmatrix}. \tag{22}$$

Calculons maintenant, pour la comparaison, la matrice des moments centrés d'ordre deux de l'estimateur θ^*. On a

$$\mathbf{E}_\theta(\theta_1^* - \theta_1)^2 = \mathbf{E}_\theta(\bar{x} - \alpha)^2 = \frac{\sigma^2}{n},$$

$$\mathbf{E}_\theta(\theta_2^* - \theta_2)^2 = \mathbf{E}_\theta(S_0^2 - \sigma^2)^2 = \frac{2\sigma^4}{n-1},$$

$$\mathbf{E}_\theta(\theta_1^* - \theta_1)(\theta_2^* - \theta_2) = 0.$$

Les deux dernières égalités s'établissent par un calcul immédiat. Considérons par exemple la deuxième d'entre elles. Il nous suffit de vérifier que

$$\mathbf{E}_\theta(\bar{x} - \alpha)S_0^2 = 0. \tag{23}$$

Mais

$$S_0^2 = \frac{1}{n - 1}\left[\sum(x_i - \alpha)^2 - (\bar{x} - \alpha)^2\right],$$

$$(\bar{x} - \alpha)S_0^2 = \frac{1}{n(n-1)}\left[\sum(x_i - \alpha)\right]\left[\sum(x_i - \alpha)^2\right] - \frac{1}{n(n-1)}(\bar{x} - \alpha)^3.$$

Comme

$$\mathbf{E}_\theta(\bar{x} - \alpha)^3 = \mathbf{E}_\theta(x_i - \alpha)^3 = \mathbf{E}_\theta(x_i - \alpha)(x_j - \alpha)^2 = 0,$$

on obtient (23).

La matrice des moments d'ordre deux de $\theta^* - \theta$ est donc égale à

$$\begin{Vmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/(n - 1) \end{Vmatrix}.$$

Il est évident que cette matrice diffère sensiblement de la matrice $(nI(\theta))^{-1}$ seulement pour les petits n.

4. Quelques conclusions. Fermons ce paragraphe en dressant le bilan des recherches réalisées dans les six derniers. Le principal objectif consistait à trouver des méthodes de construction d'estimateurs optimaux (dans un sens ou dans un autre) et à déterminer les bornes inférieures de leurs erreurs quadratiques moyennes. Quatre voies essentielles ont été dégagées.

1. Construction d'estimateurs bayésiens (si l'on dispose d'une information *a priori* sur θ) et d'estimateurs minimax.

2. Détermination des statistiques exhaustives complètes (ou minimales) S. L'estimateur $\theta_S^* = \mathbf{E}_\theta(\theta^* \mid S)$ sera efficace dans la classe K_b contenant θ^*.

3. Utilisation des estimateurs du maximum de vraisemblance dans les cas où est réalisé le critère (3) du théorème 1 (ou le critère (15) du théorème 1A). On obtient aussi des estimateurs efficaces (voire même R-efficaces) dans les classes à biais fixé.

4. Comparaison de l'erreur quadratique moyenne $\mathbf{E}_\theta(\theta^* - \theta)^2$ de l'estimateur θ^* avec la borne inférieure R définie par l'inégalité de Rao-Cramer. Si le rapport $\mathbf{E}_\theta(\theta^* - \theta)^2/R$ est proche de l'unité, l'estimateur θ^* peut être retenu. Cette approche donne lieu à des résultats assez généraux liés à la construction d'estimateurs asymptotiquement efficaces, asymptotiquement bayésiens et asymptotiquement minimax.

Faisons la remarque suivante. Dans toutes les voies mentionnées plus haut, la forme de la dépendance de la distribution $\mathbf{P}_\theta$ par rapport au paramètre θ joue un rôle fondamental. Mais en pratique il n'est pas rare qu'on ait à estimer non pas le paramètre θ lui-même mais une fonction $\varphi(\theta)$. Ceci étant, il est aisé de voir (cf. exemple du schéma de Bernoulli dans (8.4), (8.5)) que l'estimateur $\varphi^* = \varphi(\theta^*)$ ne possède pas toujours les propriétés de l'estimateur θ^* (absence de biais, efficacité, etc. Restent valables les seules propriétés d'efficacité asymptotique si φ est une fonction régulière). De ce point de vue, il est naturel d'estimer dès le départ des *fonctions* $\varphi(\theta)$ du paramètre initial θ. Nous avons renoncé à cette approche en raison de la notable complication de nombreux résultats fondamentaux. Par ailleurs, si φ est une application bijective, l'estimation de $\varphi(\theta)$ se ramène à un problème déjà étudié, moyennant une « reparamétrisation », c'est-à-dire l'introduction d'un nouveau paramètre $\gamma = \varphi(\theta)$ auquel correspondra la famille de distributions $\mathbf{G}_\gamma = \mathbf{P}_{\varphi^{-1}(\gamma)}$.

§ 17*. Propriétés de la quantité d'information de Fisher

Nous avons déjà vu et nous aurons encore l'occasion de nous en assurer dans la suite que la quantité d'information de Fisher joue un rôle important en statistique mathématique. Etudions quelques-unes de ses propriétés.

1. Cas scalaire. La quantité d'information de Fisher

$$I(\theta) = \int \frac{(f'_\theta(x))^2}{f_\theta(x)}\,\mu(dx) = \mathbf{E}_\theta[l'(x_1, \theta)]^2$$

a fait son apparition dans le paragraphe précédent. La quantité

$$I^X(\theta) = \mathbf{E}_\theta[L'(X, \theta)]^2$$

est traitée généralement comme la *mesure de la quantité d'information contenue dans un échantillon X sur le paramètre θ*. Dans le théorème 16.1 nous avons prouvé l'*additivité* de la quantité d'information : $I^X(\theta) = nI(\theta)$, c'est-à-dire que $I^X(\theta)$ est égale à la somme des quantités d'information $I^{x_i}(\theta) = \mathbf{E}_\theta[l'(x_i, \theta)]^2 = I(\theta)$ contenues dans les observations indépendantes $x_1, \ldots, x_n$.

Prouvons encore une propriété de la quantité d'information de Fisher. Soit $S = S(X)$ une statistique à valeurs dans R^l et soit $g_\theta(s)$ la densité de la distribution de S induite par la distribution $\mathbf{P}_\theta$ dans $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ par rapport à une mesure λ dans $(R^l, \mathfrak{B}^l)$. Conformément aux notations précédentes, la quantité

$$I^S(\theta) = \mathbf{E}_\theta[(\log g_\theta(S))']^2$$

sera appelée *quantité d'information contenue dans la statistique S sur le paramètre θ*.

Signalons que la valeur $I^S(\theta)$ est indépendante de la mesure λ. En effet, soit $\tilde{\lambda}$ une autre mesure et $\nu = \lambda + \tilde{\lambda}$. Alors λ et $\tilde{\lambda}$ sont absolument continues par rapport à ν, et la densité $g^*_\theta(s)$ de la distribution de S par rapport à la mesure ν est égale à

$$g^*_\theta(s) = g_\theta(s)\,\frac{d\lambda}{d\nu} = \tilde{g}_\theta(s)\,\frac{d\tilde{\lambda}}{d\nu},$$

où $\tilde{g}_\theta$ est la densité par rapport à $\tilde{\lambda}$. Puisque $\frac{d\lambda}{d\nu}$ et $\frac{d\tilde{\lambda}}{d\nu}$ sont indépendantes de θ, les dérivées des logarithmes de ces trois expressions seront confondues.

THÉORÈME 1. *Supposons que les densités $f_\theta(x)$ et $g_\theta(s)$ vérifient les conditions (R). Alors*

$$I^S(\theta) \leqslant I^X(\theta). \tag{1}$$

L'égalité est réalisée si et seulement si S est une statistique exhaustive.

DÉMONSTRATION. Pour tout $B \in \mathfrak{B}^l$ désignons par $S^{-1}(B) \in \mathfrak{B}^n_{\mathscr{X}}$ l'ensemble des $x \in \mathscr{X}^n$ tels que $S(x) \in B$. Par définition de l'espérance mathématique

conditionnelle on a alors

$$\int_{S^{-1}(B)} L'(x, \theta)\mathbf{P}_\theta(dx) = \mathbf{E}_\theta[L'(X, \theta) ; X \in S^{-1}(B)] =$$

$$= \mathbf{E}_\theta[\mathbf{E}_\theta(L'(X, \theta) \mid S) ; S \in B]. \tag{2}$$

Par ailleurs,

$$\int_{S^{-1}(B)} L'(x, \theta)\mathbf{P}_\theta(dx) = \frac{\partial}{\partial\theta} \int_{S^{-1}(B)} f_\theta(x)\mu^n(dx) = \frac{\partial}{\partial\theta}\int_B g_\theta(s)\lambda(ds) =$$

$$= \int_B \frac{\partial}{\partial\theta} g_\theta(s)\lambda(ds) = \mathbf{E}_\theta[(\log g_\theta(S))' ; S \in B]. \tag{3}$$

En comparant (2) et (3), on voit que

$$\mathbf{E}_\theta(L'(X, \theta) \mid S) = (\log g_\theta(S))' \tag{4}$$

$[\mathbf{P}_\theta]$-presque partout. On a d'autre part

$$0 \leqslant \mathbf{E}_\theta[L'(X, \theta) - (\log g_\theta(S))']^2 =$$

$$= I^X(\theta) + I^S(\theta) - 2\mathbf{E}_\theta L'(X, \theta)(\log g_\theta(S))',$$

où en vertu de (4)

$$\mathbf{E}_\theta L'(X, \theta)(\log g_\theta(S))' =$$

$$= \mathbf{E}_\theta[(\log g_\theta(S))'\mathbf{E}_\theta(L'(X, \theta) \mid S)] = \mathbf{E}_\theta[(\log g_\theta(S))']^2 = I^S(\theta),$$

ce qui prouve l'inégalité (1).

Supposons maintenant que S est une statistique exhaustive pour θ. On a alors

$$f_\theta(X) = \psi(S, \theta)h(X). \tag{5}$$

Prenons pour λ la mesure

$$\lambda(B) = \int_{S^{-1}(B)} h(x)\mu^n(dx).$$

Comme prouvé au lemme 15.1, la distribution de S sera alors absolument continue par rapport à λ et admettra une densité $g_\theta(s) = \psi(s, \theta)$. De là on déduit, compte tenu de (5), que

$$I^X(\theta) = \mathbf{E}_\theta[L'(X, \theta)]^2 = \mathbf{E}_\theta[(\log\psi(S, \theta))']^2 = I^S(\theta).$$

Prouvons maintenant que si l'égalité $I^X(\theta) = I^S(\theta)$ est réalisée pour tous les θ, la statistique S est exhaustive. En effet, $I^X(\theta)$ est la variance de $L'(X, \theta)$, de sorte que

$$I^X(\theta) = \mathbf{E}_\theta[L'(X, \theta) - \mathbf{E}_\theta(L'(X, \theta) \mid S)]^2 + \mathbf{E}_\theta[\mathbf{E}_\theta(L'(X, \theta) \mid S)]^2. \tag{6}$$

Mais en vertu de (4) le dernier terme est égal à

$$\mathsf{E}_\theta[(\log g_\theta(S))']^2 = I^S(\theta).$$

Puisque $I^X(\theta) = I^S(\theta)$, dans (6) on a $[\mathbf{P}_\theta]$-presque partout pour tous les θ

$$L'(X, \theta) - \mathsf{E}_\theta(L'(X, \theta) \mid S) = 0.$$

Autrement dit, $L'(X, \theta)$ est mesurable par rapport à $\sigma(S)$ et par suite, il existe une fonction mesurable $\varphi(S, \theta)$ telle que

$$L'(X, \theta) = \varphi(S, \theta), \; L(X, \theta) = \Phi(S, \theta) + h_1(X),$$
$$f_\theta(X) = \exp[\Phi(S, \theta) + h_1(X)]. \blacktriangleleft$$

Nous avons déjà signalé que les statistiques exhaustives étaient les seules statistiques à réduire les données empiriques sans perte d'information sur le paramètre θ. Le théorème 1 confère à cette proposition une signification rigoureuse dans le cas de la quantité d'information de Fisher.

EXEMPLE 1. Soit $X \in \mathbf{B}_p$. On a ici

$$f_p(x) = p^x(1 - p)^{1-x},$$

où x est égale à 0 ou à 1, $f_p(x)$ est la densité par rapport à la mesure cardinale. Donc

$$l(x, p) = x \ln p + (1 - x) \ln(1 - p),$$
$$l'(x, p) = \frac{x}{p} - \frac{1 - x}{1 - p},$$

$$I(p) = \mathsf{E}_p[l'(x_1, p)]^2 = p\left(\frac{1}{p}\right)^2 + (1 - p)\left(\frac{1}{1 - p}\right)^2 = \frac{1}{p(1 - p)}.$$

Par conséquent, la quantité d'information contenue dans une seule observation dans le schéma de Bernoulli est égale à $(p(1-p))^{-1}$ et atteint son minimum pour $p = 1/2$.

La quantité d'information contenue dans l'échantillon tout entier est égale à $n/(p(1-p))$. Désignons maintenant par ν le nombre de « succès » dans l'échantillon X (le nombre d'unités) et trouvons la quantité d'information contenue dans cette observation. Les densités (par rapport à la mesure cardinale) de ν seront égales à

$$g_p(x) = C_n^x p^x(1 - p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

de sorte que

$$\log g_p(x) = x \log p + (n - x)\log(1 - p) + \log C_n^x,$$

$$I^{\nu}(p) = \mathbf{E}_p[(\log g_p(\nu))']^2 = \sum_{x=0}^{n} C_n^x p^x (1-p)^{n-x}\left(\frac{x}{p} - \frac{n-x}{1-p}\right)^2 =$$

$$= \sum_{x=0}^{n} C_n^x p^x (1-p)^{n-x} \frac{(x-np)^2}{(p(1-p))^2} = \frac{1}{(p(1-p))^2}\mathbf{V}\nu = \frac{n}{p(1-p)}.$$

Cette égalité est conforme au théorème 1.

Nous proposons au lecteur de trouver à titre d'exercice les quantités d'informations contenues dans des échantillons dont les distributions dépendent d'un paramètre scalaire (cf. § 2).

2. Cas vectoriel. Supposons maintenant que $\theta \in R^k$, $k > 1$. Dans ce cas nous aurons affaire à la *matrice d'information de Fisher* de l'observation x_1

$$I(\theta) = \| I_{ij}(\theta) \|,\ I_{ij}(\theta) = \mathbf{E}_\theta \frac{\partial}{\partial \theta_i} l(x_1, \theta) \frac{\partial}{\partial \theta_j} l(x_1, \theta),$$

où l'on admet évidemment que la fonction $f_\theta(x)$ est dérivable.

Si l'on pose

$$\varphi(x, \theta) = (\varphi_1(x, \theta), \ldots, \varphi_k(x, \theta)) =$$

$$= 2(\sqrt{f_\theta(x)})' = \frac{1}{\sqrt{f_\theta(x)}}\left(\frac{\partial f_\theta(x)}{\partial \theta_1}, \ldots, \frac{\partial f_\theta(x)}{\partial \theta_k}\right),$$

on peut mettre la matrice $I(\theta)$ sous la forme

$$I(\theta) = \int_{\mathscr{X}} \varphi^T(x, \theta)\varphi(x, \theta)\mu(dx).$$

Nous avons déjà établi au § 16 que, de même que dans le cas scalaire, la quantité d'information de Fisher est additive, c'est-à-dire que la matrice d'information de Fisher de l'échantillon X est égale à la somme des matrices d'information des diverses observations. Si l'on pose

$$I^X(\theta) = \| I_{ij}^X(\theta) \|,\ I_{ij}^X(\theta) = \mathbf{E}_\theta \frac{\partial}{\partial \theta_i} L(X, \theta) \frac{\partial}{\partial \theta_j} L(X, \theta),$$

alors $I^X(\theta) = nI(\theta)$.

Le théorème 1 reste entièrement en vigueur. Supposons que $g_\theta(s)$ est la densité d'une statistique $S = S(X)$ à valeurs dans R^l par rapport à une mesure λ. Posons

$$I^S(\theta) = \| I_{ij}^S(\theta) \|,\ I_{ij}^S(\theta) = \mathbf{E}_\theta \frac{\partial}{\partial \theta_i} \log g_\theta(S) \frac{\partial}{\partial \theta_j} \log g_\theta(S).$$

Ceci est la matrice d'information de l'observation S.

THÉORÈME 1A. *Si les densités $f_\theta(x)$ et $g_\theta(s)$ satisfont les conditions (R) du* § 16, *alors*

$$I^S(\theta) \leqslant I^X(\theta), \tag{7}$$

c'est-à-dire que la matrice $I^X(\theta) - I^S(\theta)$ est semi-définie positive. Dans (7) *l'égalité est réalisée si et seulement si S est une statistique exhaustive.*

DÉMONSTRATION. Elle est entièrement calquée sur celle du théorème 1. Nous l'omettrons pour abréger l'exposé. Le lecteur intéressé pourra la trouver par exemple dans [42] et [91].

EXEMPLE 2. Au § 16 nous avons déjà calculé la matrice d'information pour la distribution normale. Calculons-la maintenant pour la famille à deux paramètres de distributions de densité

$$f_\theta(x) = \frac{1}{\sigma} f\left(\frac{x-\alpha}{\sigma}\right).$$

Ici $\theta = (\alpha, \sigma)$ et f est une fonction dérivable donnée telle qu'existent les intégrales

$$I_i = \int x^i \frac{(f'(x))^2}{f(x)}\, dx = \mathbf{E}_{(0,\,1)} x_1^i (l'(x_1))^2,\ i = 0,\ 1,\ 2,$$

où $l(x) = \log f(x)$, le symbole « prime » désigne la dérivation ordinaire et les paramètres α et σ sont respectivement les paramètres de translation et d'échelle de la distribution de densité $f(x)$. Nous connaissons donc la forme de la distribution à une transformation linéaire près de l'argument. Les paramètres α et σ de la distribution normale Φ_{α,σ^2} sont visiblement des paramètres de translation et d'échelle. Le paramètre α de la distribution gamma pour λ fixe est paramètre d'échelle au même titre que θ l'est pour la distribution $\mathbf{U}_{0,\theta}$.

On a

$$l(x, \theta) = \log f_\theta(x) = -\log\sigma + l\left(\frac{x-\alpha}{\sigma}\right),$$

$$\frac{\partial l(x,\theta)}{\partial\alpha} = -\frac{1}{\sigma} l'\left(\frac{x-\alpha}{\sigma}\right),$$

$$\frac{\partial l(x,\theta)}{\partial\sigma} = -\frac{1}{\sigma} - \frac{(x-\alpha)}{\sigma^2} l'\left(\frac{x-\alpha}{\sigma}\right),$$

d'où

$$I_{11}(\theta) = \frac{1}{\sigma^2}\mathbf{E}_\theta\left[l'\left(\frac{x_1-\alpha}{\sigma}\right)\right]^2 = \frac{1}{\sigma^2}\int \frac{\left[f'\left(\frac{x-\alpha}{\sigma}\right)\right]^2}{\sigma f\left(\frac{x-\alpha}{\sigma}\right)}\, dx = \frac{1}{\sigma^2} I_0,$$

$$I_{12}(\theta)=\frac{1}{\sigma^2}\,\mathsf{E}_\theta l'\left(\frac{x_1-\alpha}{\sigma}\right)\left[1+\frac{x_1-\alpha}{\sigma}\,l'\left(\frac{x_1-\alpha}{\sigma}\right)\right]=\frac{1}{\sigma^2}\,I_1,$$

$$I_{22}(\theta)=\frac{1}{\sigma^2}\,\mathsf{E}_\theta\left[1+\frac{x_1-\alpha}{\sigma}\,l'\left(\frac{x_1-\alpha}{\sigma}\right)\right]^2=\frac{1}{\sigma^2}\,[I_2-1],$$

puisque

$$2\int\frac{x-\alpha}{\sigma}\,f'\left(\frac{x-\alpha}{\sigma}\right)\frac{dx}{\sigma}=-2\int f(x)dx=-2.$$

Donc

$$I(\theta)=\frac{1}{\sigma^2}\begin{Vmatrix}I_0 & I_1\\ I_1 & I_2-1\end{Vmatrix}.$$

Si f est une fonction symétrique, il est évident que $I_1=0$.

La dégénérescence de la matrice $I(\theta)$ exprime que son déterminant est nul ou, ce qui revient au même, que

$$[\mathsf{E}_{(0,\,1)}l'(x_1)(1+x_1l'(x_1))]^2=\mathsf{E}_{(0,\,1)}(l'(x_1))^2\mathsf{E}_{(0,\,1)}(1+x_1l'(x_1))^2.$$

Ceci n'est possible que si ou bien $1+xl'(x)=cl'(x)$ pour un c quelconque, ou bien $l'(x)=0$. La première égalité entraîne

$$l(x)=-\ln(x-c)+c_1,\quad f(x)=\frac{e^{c_1}}{x-c}.$$

Il est évident qu'une telle fonction $f(x)$ ne peut être densité d'une distribution. On traite de façon analogue le cas $l'(x)=0$. Donc, $I(\theta)$ est définie positive.

Pour la famille normale $\{\Phi_{\alpha,\sigma^2}\}$, $\theta=(\alpha,\sigma)$, on a en particulier

$$I(\theta)=\frac{1}{\sigma^2}\begin{Vmatrix}1 & 0\\ 0 & 2\end{Vmatrix},$$

puisque dans ce cas $l(x)=-x^2/2-\ln\sqrt{2\pi}$, $l'(x)=-x$, $I_0=\mathsf{E}_{(0,\,1)}x_1^2=1$, $I_1=\mathsf{E}_{(0,\,1)}x_1^3=0$, $I_2=\mathsf{E}_{(0,\,1)}x_1^4=3$. On aurait pu aboutir au même résultat en considérant l'exemple 16.4 et en se servant du numéro 3 ci-dessous où l'on étudie le comportement de la matrice d'information sous l'effet d'un changement de paramètre (dans l'exemple 16.4 on a $\theta=(\alpha,\sigma^2)$ et non pas $\theta=(\alpha,\sigma)$). Nous proposons au lecteur de s'assurer maintenant qu'en vertu du théorème 1A la statistique $\left(\bar{x},\sum x_i^2\right)$ admet la matrice d'information

$$I^S(\theta)=\frac{n}{\sigma^2}\begin{Vmatrix}1 & 0\\ 0 & 2\end{Vmatrix}=nI(\theta).$$

3. Matrice de Fisher et changement de paramètre. Voyons comment se comporte la matrice d'information sous l'effet d'un changement de paramètre. Posons $\theta=v(\beta)$, $\beta\in R^k$, où v est une fonction vectorielle dérivable, et

considérons la famille paramétrique $\mathbf{P}_\beta^{(1)} = \mathbf{P}_{v(\beta)}$. Pour trouver la matrice d'information $J(\beta)$ de cette famille, nous devons calculer les dérivées

$$\frac{\partial}{\partial \beta_j} l(\mathrm{x}_1, v(\beta)) = \sum_{i=1}^{k} \frac{\partial}{\partial \theta_i} l(\mathrm{x}_1, v(\beta)) \frac{\partial v_i(\theta)}{\partial \beta_j}. \tag{8}$$

Si l'on pose $V = \left\| \frac{\partial v_i(\beta)}{\partial \beta_j} \right\|$, $i, j, = 1, \ldots, k$, on trouve que le vecteur $l'_\beta(\mathrm{x}_1, v(\beta))$ des dérivées dans (8) se représente sous la forme $l'_\theta(\mathrm{x}_1, v(\beta))V$, de sorte que

$$J(\beta) = \mathbf{E}_\beta(l'_\theta(\mathrm{x}_1, v(\beta))V)^T(l'_\theta(\mathrm{x}_1, v(\beta))V) = V^T I(v(\beta))V.$$

En particulier, si $\theta = \beta C$, $C = \| c_{ij} \|$, $i, j = 1, \ldots, k$, alors $V = C^T$ et

$$J(\beta) = CI(\theta)C^T. \tag{9}$$

A noter que l'équation paramétrique de l'ellipsoïde

$$(\theta - \theta_1)I(\theta)(\theta - \theta_1)^T < c \tag{10}$$

est invariante par une transformation linéaire inversible C sur θ. Plus exactement, si l'on pose $\theta = \beta C$, l'inéquation (10) devient

$$(\beta - \beta_1)J(\beta)(\beta - \beta_1)^T < c,$$

où $\beta_1 = \theta_1 C^{-1}$. On obtient immédiatement cette inéquation en portant $\theta = \beta C$ dans (10) et en se servant de (9).

§ 18*. Estimateurs des paramètres de translation et d'échelle. Estimateurs efficaces équivariants

Nous avons vu aux §§ 12 à 16 et nous verrons dans la suite combien la notion de statistique exhaustive est utile en général et dans la construction des estimateurs efficaces en particulier. Tout ce qui est rattaché à l'utilisation des statistiques exhaustives pourrait être appelé *principe d'exhaustivité.*

Pour construire des estimateurs efficaces nous avons combiné le principe d'exhaustivité à un autre principe : le *principe d'absence de biais.* Ce dernier consiste à mettre en évidence une classe d'estimateurs de biais fixé et en particulier de biais nul. Sans fixer le biais il est impossible de construire des estimateurs efficaces.

Dans ce paragraphe et les suivants, ainsi que dans le chapitre 3, nous étudierons un autre principe important de statistique mathématique : le *principe d'invariance.*

Ces principes poursuivent le même objectif : leur introduction permet de restreindre de façon naturelle les classes des estimateurs considérés de

telle sorte qu'il soit possible de trouver des estimateurs efficaces dans les restrictions obtenues.

1. Estimateurs des paramètres de translation et d'échelle. L'estimation du paramètre de translation est le problème qui consiste à estimer le paramètre α dans une famille de distributions $\{\mathbf{P}_\alpha\}$ telle que

$$\mathbf{P}_\alpha(A) = \mathbf{P}(A - \alpha).$$

Ici $\mathbf{P}$ est une distribution fixe, $A - \alpha = \{x : x + \alpha \in A\}$ et l'on admet que l'ensemble Θ est de même nature que $\mathscr{X}$. Si $\mathscr{X} = R^m$, on peut de toute évidence considérer des paramètres θ de « moindre dimension », par exemple des paramètres scalaires, mais il faut alors fixer le sens (le vecteur $e \in \mathscr{X}$) de la translation et étudier $\mathbf{P}_\alpha(A) = \mathbf{P}(A + \alpha e)$. Pour fixer les idées on ne traitera que le premier cas et l'on admettra que $\Theta = \mathscr{X} = R^m$.

A noter que la distribution $\mathbf{P}_\alpha$ de $x_i + c$ ($c \in R^m$) est confondue avec la distribution $\mathbf{P}_{\alpha+c}$ de x_i, c'est-à-dire qu'une c-translation des observations nous conduit à un échantillon de distribution $\mathbf{P}_{\alpha+c}$. Il est donc naturel de n'étudier que les estimateurs $\alpha^* = \alpha^*(X)$ du paramètre α tels que

$$\alpha^*(X + c) = \alpha^*(X) + c. \tag{1}$$

Ici et dans la suite, $X + c$ représente le vecteur de coordonnées $(x_1 + c, \ldots, x_n + c)$. La violation de cette égalité exprime que l'estimateur α^* dépend de l'origine du système de référence, c'est-à-dire de l'origine des coordonnées de l'espace $\mathscr{X} = R^m$.

On procède de même quand on estime le paramètre d'échelle σ de la famille $\{\mathbf{P}_\sigma\}$ telle que

$$\mathbf{P}_\sigma(A) = \mathbf{P}(A \mid \sigma), \quad \sigma \in]0, \infty[.$$

On admet que σ est scalaire, bien que l'on puisse envisager aussi le cas matriciel. Dans ce cas la distribution $\mathbf{P}_\sigma$ de $x_i c$ est confondue avec la distribution $\mathbf{P}_{\sigma c}$ de x_i, c'est-à-dire que la multiplication des observations par c conduit à un échantillon de distribution $\mathbf{P}_{\sigma c}$. On peut se borner donc à étudier des estimateurs tels que

$$\sigma^*(Xc) = c\sigma^*(X), \tag{2}$$

où $Xc = (x_1 c, \ldots x_n c)$, puisque si les observations sont multipliées par c il en est de même du paramètre d'échelle.

Le lecteur établira sans peine les propositions suivantes.

Si une famille $\mathbf{P}_\theta$ vérifie la condition (A_μ), le paramètre θ sera paramètre de translation (resp. d'échelle) si et seulement si

$$f_\theta(x) = f(x - \theta) \quad \left(\text{resp. } f_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)\right).$$

Si $\mathscr{X}=R=\Theta$, $X\in\mathbf{P}_\alpha$ et α est un paramètre de translation, alors $Y=e^X=(e^{x_1}, \ldots, e^{x_n})\in\mathbf{Q}_\sigma$, où le paramètre $\sigma=e^\alpha$ est un paramètre d'échelle pour les distributions $\mathbf{Q}_\sigma$. Ceci résulte directement du fait que la densité de $y_1=e^{x_1}$ est égale à (cf. [11])

$$\frac{1}{y}f(\ln y - \alpha) = \frac{1}{\sigma}\left[\frac{\sigma}{y}f\left(\ln\frac{y}{\sigma}\right)\right].$$

Réciproquement, si $\mathscr{X}=]0, \infty[=\Theta$, $X\in\mathbf{P}_\sigma$ et σ est un paramètre d'échelle, alors $Y=\ln X=(\ln x_1, \ldots, \ln x_n)\in\mathbf{Q}_\sigma$, où $\alpha=\ln\sigma$ est le paramètre de translation des distributions $\mathbf{Q}_\alpha$.

On peut envisager l'estimation simultanée des paramètres inconnues α et σ dans le cas où $\mathbf{P}_{\alpha,\sigma}(A)=\mathbf{P}\left(\dfrac{A-\alpha}{\sigma}\right)$. Dans ces conditions, pour estimateur de σ il est naturel de considérer des fonctions telles que

$$\alpha^*(X+c) = \alpha^*(X),\ \sigma^*(Xc) = c\sigma^*(X). \tag{3}$$

Les estimateurs vérifiant les conditions (1), (2) et (3) des exemples ci-dessus s'appellent *estimateurs équivariants* (la définition générale est donnée au § 19). L'introduction de tels estimateurs a pour but de restreindre la classe des estimateurs considérés afin de simplifier la recherche des estimateurs optimaux. Ainsi, au § 8 nous avons établi qu'il était impossible de déterminer les estimateurs uniformément (c'est-à-dire pour tous les θ) les meilleurs dans la classe de tous les estimateurs. Or il se trouve que la classe des estimateurs équivariants contient des estimateurs uniformément les meilleurs qui peuvent être, dans bien des cas, déterminés sous une forme explicite. Nous nous proposons d'illustrer ce fait sur l'exemple des estimations des paramètres de translation et d'échelle.

2. Estimateur efficace du paramètre de translation dans la classe des estimateurs équivariants. On admettra ici qu'est réalisée la condition (A_μ), donc que $f_\alpha(x)=f(x-\alpha)$ et que μ est la mesure de Lebesgue.

Désignons par S_0 la statistique

$$S_0 = S_0(X) = (x_2 - x_1, \ldots, x_n - x_1)$$

qui est visiblement invariante par une translation : $S_0(X+c)=S_0(X)$. Désignons par K_E la classe des estimateurs équivariants α^*, c'est-à-dire des estimateurs vérifiant (1), et par $|\alpha|^2$, le carré de la norme euclidienne de $\alpha\in R^m$.

THÉORÈME 1. *Soit* $\alpha^*=\alpha^*(X)$ *un estimateur équivariant dont* $\mathbf{E}_0\alpha^*$ *est finie. Alors l'estimateur*

$$\alpha_0^* = \alpha^* - \mathbf{E}_0(\alpha^* \mid S_0) \tag{4}$$

est indépendant du choix de α^* et constitue l'unique estimateur efficace de K_E, c'est-à-dire que $\mathbf{E}_\alpha|\alpha_0^* - \alpha|^2 = \min_{\alpha^* \in K_E} \mathbf{E}_\alpha|\alpha^* - \alpha|^2$ *pour tous les* α *et* $\mathbf{E}_\alpha|\alpha^* - \alpha|^2 = \mathbf{E}_\alpha|\alpha_0^* - \alpha|^2$ *si seulement* $\mathbf{E}_0(\alpha^* \mid S_0) = 0$ *presque partout. L'estimateur* α_0^* *peut être mis sous la forme*

$$\alpha_0^* = \frac{\int u f_u(X) du}{\int f_u(X) du} = \frac{\int u f(X - u) du}{\int f(X - u) du}. \tag{5}$$

L'estimateur α_0^* s'appelle *estimateur de Pitman*. Il est aisé de voir sur (4) qu'il est équivariant et sans biais. L'équivariance résulte de celle de α^* et de l'invariance, par une translation, de la fonction $V(S_0) = \mathbf{E}_0(\alpha^* \mid S_0)$ qui dépend uniquement de S_0. L'absence de biais résulte des égalités

$$\mathbf{E}_\alpha \alpha_0^* = \alpha + \mathbf{E}_\alpha \alpha^*(X - \alpha) - \mathbf{E}_\alpha V(S_0), \tag{6}$$

où $\mathbf{E}_\alpha V(S_0) = \mathbf{E}_0 V(S_0)$, $\mathbf{E}_\alpha \alpha^*(X - \alpha) = \mathbf{E}_0 \alpha^*(X)$. La dernière égalité découle du fait que $X - \alpha \in \mathbf{P}_0$, si $X \in \mathbf{P}_\alpha$. La somme des deux derniers termes de (6) est donc égale à

$$\mathbf{E}_0 \alpha^* - \mathbf{E}_0[\mathbf{E}_0(\alpha^* \mid S_0)] = 0 \; ; \; \mathbf{E}_\alpha \alpha_0^* = \alpha.$$

Etablissons préalablement la proposition auxiliaire suivante.

LEMME 1. *Soit $X \in \mathbf{P}_0$. L'espérance mathématique conditionnelle par rapport à S_0 de toute statistique $S = S(X)$ d'espérance mathématique $\mathbf{E}_0|S|$ finie est égale à*

$$\mathbf{E}_0(S \mid S_0) = S_1(X) \equiv \frac{\int S(X - u) f_u(X) du}{\int f_u(X) du}. \tag{7}$$

DÉMONSTRATION. Toutes les fonctions figurant sous les signes d'intégration de (7) sont des fonctions de $X - u$. Si l'on fait donc le changement $x_1 - u = v$, on obtiendra des fonctions de $(v, x_2 - x_1 + v, \ldots, x_n - x_1 + v)$. Ceci exprime que le second membre de (7) ne dépend que de S_0. En vertu des propriétés de l'espérance mathématique conditionnelle, pour prouver le lemme, il nous suffit de montrer que pour tout $A \in \sigma(S_0)$

$$\mathbf{E}_0(S_1 \; ; A) = \mathbf{E}_0(S \; ; A). \tag{8}$$

Soit $Z = Z(S_0)$ une statistique $\sigma(S_0)$-mesurable bornée. On a alors

$$\mathbf{E}_0 Z S_1 = \int_{\mathscr{X}^n} \frac{Z(S_0) \int_\Theta S(x - u) f_u(x) du}{\int_\Theta f_u(x) du} f(x) dx =$$

$$= \int_\Theta \int_{\mathscr{X}^n} \frac{Z(S_0) S(x - u) f(x - u) f(x)}{\int_\Theta f(x - v) dv} dx\, du.$$

Le changement $x-u\to x$ dans l'intégrale intérieure nous donne ($S_0(x)$ est invariante par ce changement)

$$\int_{\Theta}\int_{\mathscr{X}^n} \frac{Z(S_0)S(x)f(x)f(x+u)}{\int_{\Theta} f(x+u-v)dv}\,dx\,du = \int_{\mathscr{X}^n} Z(S_0)S(x)f(x)dx = \mathbf{E}_0 ZS.$$

Ce qui prouve (8). Les deux changements d'ordre d'intégration sont licites, puisque S est absolument intégrable et Z bornée. ◄

DÉMONSTRATION du théorème 1. Remarquons tout d'abord que si α^* est un estimateur équivariant, la quantité $\mathbf{E}_\alpha\,|\,\alpha^*-\alpha\,|^2$ est indépendante de α. En effet

$$\mathbf{E}_\alpha\,|\,\alpha^*(X)-\alpha\,|^2 = \mathbf{E}_\alpha\,|\,\alpha^*(X-\alpha)\,|^2 = \mathbf{E}_0\,|\,\alpha^*(X)\,|^2.$$

Pour trouver un estimateur équivariant uniformément optimal, il faut donc trouver un estimateur α^* minimisant $\mathbf{E}_0\,|\,\alpha^*\,|^2$.

Soit α^* un estimateur équivariant quelconque de α. Les propriétés de l'espérance mathématique conditionnelle nous donnent

$$\mathbf{E}_0\,|\,\alpha^*\,|^2 = \mathbf{E}_0\,|\,\alpha^* - \mathbf{E}_0(\alpha^*\,|\,S_0)\,|^2 + \mathbf{E}_0\,|\,\mathbf{E}_0(\alpha^*\,|\,S_0)\,|^2 \geqslant \\ \geqslant \mathbf{E}_0\,|\,\alpha^*-\mathbf{E}_0(\alpha^*\,|\,S_0)\,|^2. \quad (9)$$

Il reste à remarquer qu'en vertu du lemme 1, l'estimateur $\alpha_0^* = \alpha^* - \mathbf{E}_0(\alpha^*\,|\,S_0)$ est égal à (5) et ne dépend pas du choix de α^*. Il est évident que dans (9) l'égalité est possible si et seulement si $\mathbf{E}_0(\alpha^*\,|\,S_0)=0$ presque partout. ◄

De la démonstration du théorème il ressort que la statistique $S_0=(x_2-x_1,\ldots,x_n-x_1)$ qui est invariante par une translation joue un rôle particulier dans la construction d'un estimateur équivariant optimal. L'invariance de la statistique est une propriété qui dans un certain sens est contraire à l'exhaustivité, quant à la construction de l'estimateur $\theta_0^*=\theta^*-$ $-\mathbf{E}_0(\theta^*\,|\,S_0)$ qui vise à améliorer θ^*, elle est dans un certain sens aussi contraire à l'approche qui consistait à construire l'estimateur $\theta_S^*=\mathbf{E}_\theta(\theta^*\,|\,S)$ pour améliorer θ^* à l'aide de la statistique exhaustive S. Ces deux approches sont contraires en ce sens que la statistique exhaustive contient toute l'information sur θ, alors que la statistique invariante n'en contient aucune. Pour trouver les meilleurs estimateurs nous avons cherché les statistiques exhaustives minimales ; ici il nous faudra trouver des statistiques invariantes maximales (telle est la statistique S_0). L'estimateur θ_S^* est la « projection » de θ^* sur S, alors que l'estimateur θ_0^* s'obtient en soustrayant de θ^* sa « projection » sur S_0.

En définitive, les résultats obtenus par ces deux approches sont souvent confondus comme le montrent les deux exemples suivants.

EXEMPLE 1. Soient $\mathscr{X}=R$, $X\in\Phi_{\alpha,\,1}$. Alors

$$f_\alpha(X) = \frac{1}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\sum(x_i - \alpha)^2\right\} =$$
$$= \frac{1}{\sqrt{n}(2\pi)^{\frac{n-1}{2}}} \exp\left\{-\frac{1}{2}\sum(x_i - \bar{x})^2\right\}\cdot\sqrt{\frac{n}{2\pi}}\, e^{-\frac{n}{2}(\alpha-\bar{x})^2}$$

Le second facteur traité comme une fonction de α est la densité d'une loi normale de paramètres $(\bar{x}, 1/n)$. Le premier facteur étant indépendant de α, on peut le simplifier dans (5) et l'estimateur de Pitman sera égal à $\alpha^* = \bar{x}$. On obtient le même résultat pour le cas vectoriel.

EXEMPLE 2. Soient $\mathscr{X}=R$, $X \in U_{\theta,\,1+\theta}$. Alors

$$f_\theta(X) = \begin{cases} 1 & \text{pour } x_{(n)} - 1 \leqslant \theta \leqslant x_{(1)}, \\ 0 & \text{sinon.} \end{cases}$$

Donc,

$$\theta^* = \int_{x_{(n)}-1}^{x_{(1)}} u\, du \Big/ (x_{(1)} - x_{(n)} + 1) = \frac{1}{2}(x_{(1)} + x_{(n)} - 1).$$

Nous voyons par conséquent que dans la classe K_E des estimateurs équivariants, on peut construire des estimateurs efficaces sous forme explicite sans poser de conditions sur la dérivabilité de $f_\theta(x)$, l'efficacité revêtant un caractère exact et non asymptotique.

3. Minimaximalité de l'estimateur de Pitman. Portons notre attention sur la forme de l'estimateur de Pitman. En gros, c'est un estimateur bayésien pour une distribution *a priori* « uniforme sur l'axe tout entier ». Formulons cette proposition avec plus de rigueur, puisque la distribution mentionnée n'existe pas. Supposons que $\mathscr{X}=R$ et que $Q^{(N)}$ est une distribution uniforme sur $[-N, N]$, c'est-à-dire une distribution de densité

$$q^{(N)}(t) = \begin{cases} (2N)^{-1}, & |t| \leqslant N, \\ 0, & |t| > N. \end{cases}$$

L'estimateur bayésien correspondant à $Q^{(N)}$ sera égal à

$$\alpha^*_{Q^{(N)}} = \frac{\int u q^{(N)}(u) f_u(X) du}{\int q^{(N)}(u) f_u(X) du} = \int_{-N}^{N} u f_u(X) du \Big/ \int_{-N}^{N} f_u(X) du.$$

Il est évident que pour tous les X, l'estimateur de Pitman α_0^* est la limite $\alpha_0^* = \lim_{N\to\infty} \alpha^*_{Q^{(N)}}$. Ceci nous suggère la convergence simultanée des moments d'ordre deux :

$$E_\alpha(\alpha^*_{Q^{(N)}} - \alpha)^2 \to E_\alpha(\alpha_0^* - \alpha)^2.$$

Il s'avère que cette convergence a bien lieu et qu'elle est uniforme par rapport à α dans le domaine $|\alpha| \leqslant N-\sqrt{N}$. (La démonstration qui implique l'estimation de $\mathbf{E}_\alpha(\overset{\bullet}{\alpha}_0-\overset{\bullet}{\alpha}_{Q^{(N)}})^2$ est purement technique, aussi l'omettrons-nous.)

Mais dans ce cas nous pouvons utiliser le critère de minimaximalité des estimateurs du théorème 11.3 : *si un estimateur* α^* *est tel que pour tous les* α

$$\mathbf{E}_\alpha(\alpha^* - \alpha)^2 \leqslant \operatorname*{lim.sup}_{N\to\infty}\int \mathbf{E}_t(\overset{\bullet}{\alpha}_{Q^{(N)}} - t)^2\, \mathbf{Q}^{(N)}(dt), \tag{10}$$

où $\mathbf{Q}^{(N)}$ *sont des distributions a priori (pas nécessairement uniformes) et* $\overset{\bullet}{\alpha}_{Q^{(N)}}$, *les estimateurs bayésiens correspondants, alors* α^* *est un estimateur minimax.*

Dans notre cas, $m = \mathbf{E}_\alpha(\overset{\bullet}{\alpha}_0-\alpha)^2$ est indépendant de α. En vertu des propriétés de la convergence mentionnées plus haut, on obtient donc

$$\operatorname*{lim.\,sup}_{N\to\infty}\int \mathbf{E}_t(\overset{\bullet}{\alpha}_{Q^{(N)}} - t)^2\mathbf{Q}^{(N)}(dt) \geqslant$$

$$\geqslant \operatorname*{lim.\,sup}_{N\to\infty} \frac{1}{2N} \int_{|t|<N-\sqrt{N}} \mathbf{E}_t(\overset{\bullet}{\alpha}_{Q^{(N)}} - t)^2 dt \geqslant$$

$$\geqslant \operatorname*{lim.\,sup}_{N\to\infty} \frac{1}{2N}\, 2(N - \sqrt{N})(m - \epsilon) = m - \epsilon$$

pour tout $\epsilon>0$. Ceci exprime que la propriété (10) a lieu.

L'estimateur de Pitman est donc un estimateur minimax dans la classe de tous les estimateurs du paramètre de translation (qu'il soit minimax dans la classe des estimateurs équivariants résulte de toute évidence de son efficacité).

Ce qui précède peut être interprété de la manière suivante : la distribution *a priori* « la plus défavorable » (cf. § 11) du paramètre de translation est la distribution « uniforme sur l'axe tout entier ».

L'indépendance de $\mathbf{E}_\alpha(\overset{\bullet}{\alpha}_0-\alpha)^2$ par rapport à α (comparer avec le théorème 11.2) aurait pu servir également de critère de minimaximalité de l'estimateur de Pitman.

4. Sur les estimateurs optimaux du paramètre d'échelle. Nous avons déjà signalé que l'estimation du paramètre d'échelle σ pouvait être ramenée dans un certain sens à celle du paramètre de translation. Supposons pour simplifier que $\mathscr{X}=]0, \infty[=\Theta$. Si $X\in\mathbf{P}_\sigma$, $\mathbf{P}_\sigma(A)=\mathbf{P}(A \mid \sigma)$, alors $Y=\ln X=(\ln x_1, \ldots$ $\ldots, \ln x_n)\in \mathbf{P}_\alpha^{(1)}$, où $\alpha=\ln\sigma$, et la distribution $\mathbf{P}_\alpha^{(1)}$ admet une densité égale à celle de $y_1=\ln x_1$ au point y $\left(\text{la condition } (A_\mu) \text{ est remplie, } \frac{d\mathbf{P}_1(x)}{d\mu} = f(x)\right)$, c'est-à-dire (voir [11])

$$f\left(\frac{e^y}{\sigma}\right)\frac{e^y}{\sigma} = f\left(e^{y-\alpha}\right)e^{y-\alpha} = f^{(1)}(y-\alpha),$$

$$f^{(1)}(y) = f(e^y)e^y.$$

Ainsi, nous pouvons estimer le paramètre α de la meilleure façon à l'aide de l'estimateur de Pitman $\alpha^* = \alpha^*(Y)$ et poser ensuite $\sigma^*(X) = e^{\alpha^*(Y)}$. Il est immédiat de voir que $\sigma^*(X)$ sera équivariant, puisque

$$\sigma^*(cX) = e^{\alpha^*(Y+\ln c)} = e^{\alpha^*(Y)+\ln c} = c\sigma^*(X).$$

Cependant il importe de signaler ici que l'estimateur de Pitman minimise $\mathbf{E}_\alpha(\alpha^* - \alpha)^2$. Donc, l'estimateur σ^* obtenu minimisera la quantité

$$\mathbf{E}_\sigma\left(\ln\frac{\sigma^*}{\sigma}\right)^2 \tag{11}$$

et non pas la quantité $\mathbf{E}_\sigma(\sigma^* - \sigma)^2$ à laquelle nous avions ordinairement affaire. Mais en cherchant un estimateur équivariant du paramètre σ on n'a aucun intérêt à considérer l'erreur quadratique moyenne, puisque contrairement à (11) elle dépend d'une application contractante portant simultanément sur σ^* et sur σ. L'analogue de la statistique invariante S_0 sera ici la statistique $(x_2/x_1, \ldots, x_n/x_1)$. On peut évidemment considérer des erreurs autres que (11). Si par exemple l'on minimise la quantité

$$\mathbf{E}_\sigma\left(\frac{\sigma^*}{\sigma} - 1\right)^2,$$

le meilleur estimateur équivariant sera

$$\sigma^* = \frac{\int \sigma^{-n-2} f(X/\sigma)d\sigma}{\int \sigma^{-n-3} f(X/\sigma)d\sigma} \tag{12}$$

(cf. [27]).

EXEMPLE 3. *Détection d'une source de rayonnement.* Citons un exemple de problème de physique lié à l'estimation des paramètres de translation et d'échelle.

Supposons qu'une source de rayons γ est placée en un point inconnu z de l'espace. Le problème consiste à déterminer les coordonnées de z en repérant sur un détecteur plan (supposé confondu avec un plan de coordonnées) les traces du rayonnement, c'est-à-dire les traces de l'action des rayons émis par le point z sur la surface sensible du détecteur.

Ce problème aurait été grandement simplifié si l'on avait eu affaire à une source de particules chargées douées d'une haute énergie. On aurait pu alors placer l'un à la suite de l'autre deux détecteurs plans parallèles et fixer les points de passage (c'est-à-dire d'action sur la surface de l'écran) de deux particules en tout. Ceci nous aurait fourni la direction de vol de ces particules

et, partant, les coordonnées de leur point d'intersection z. Mais ce procédé est irréalisable pour le faible rayonnement γ utilisé en radioscopie et l'on ne peut se servir que d'un seul détecteur.

Les rayons γ se propagent dans une direction aléatoire qui est uniformément distribuée sur une sphère (si cette direction est définie par un point sur une sphère centrée en z).

Pour simplifier ce problème nous le traiterons dans le plan. Supposons que la source est située en un point $z=(\alpha, \sigma)$, $\sigma>0$, du plan (x, y). L'angle que fait la direction du rayonnement avec l'axe Oy est uniformément distribué sur $[0, 2\pi]$. Le détecteur est confondu avec l'axe des abscisses. Les résultats des observations seront les points x_1, x_2, ... d'impact des rayons γ sur le détecteur.

Le trait spécifique de ce problème est que la taille n de l'échantillon obtenu durant un intervalle de temps t fixe sera aléatoire : le nombre des rayons gamma émis par la source durant le temps t suit la loi de Poisson de même que le nombre des rayons gamma atteignant le détecteur, puisque chaque rayon atteint l'axe des abscisses avec une probabilité 1/2. Mais, dans notre cas, le nombre n et les observations x_1, x_2, ... sont indépendants. Nous pouvons donc envisager le nombre n d'observations obtenues et admettre qu'il est fixe (la distribution de x_i sera la même pour chaque n ainsi obtenu).

Soient données les observations $X=(\mathrm{x}_1, \ldots, \mathrm{x}_n)$. Notre problème consiste à estimer les coordonnées (α, σ). Montrons que $X\in \mathbf{K}_{\alpha,\,\sigma}$, c'est-à-dire que x_i sont distribuées suivant la loi de Cauchy de paramètres de translation α et d'échelle σ.

En effet, la distribution conditionnelle de l'angle β formé par la direction de propagation d'un rayon γ avec l'axe $(0, -y)$ sachant que ce rayon atteint le détecteur sera uniforme sur l'intervalle $[-\pi/2, \pi/2]$. Comme $(x-\alpha)/\sigma=\operatorname{tg}\beta$ (cf. fig. 2), il vient

$$\mathbf{P}_{\alpha,\,\sigma}(\mathrm{x}_1 < x) = \frac{1}{2} + \frac{1}{\pi}\operatorname{Arctg}\frac{x-\alpha}{\sigma}.$$

La densité de la distribution de x_1 sera donc égale à la densité de la distribution de Cauchy (cf. § 2)

$$k_{\alpha,\,\sigma}(x) = \frac{1}{\pi\sigma}\,\frac{1}{1+((x-\alpha)/\sigma)^2} = \frac{\sigma}{\pi(\sigma^2+(x-\alpha)^2)}.$$

Supposons maintenant que σ est connu, par exemple $\sigma=1$. Le meilleur estimateur invariant du paramètre de translation α sera alors l'estimateur de

Pitman qui est égal à la moyenne $\alpha^* = \int u\varphi(u)du$ de la distribution de densité

$$\varphi(u) = \varphi(u, X) = \frac{k_u(X)}{\int k_v(X)dv},$$

$$k_u(X) = \prod_{i=1}^{n} k_u(x_i),\ k_u(x_i) = k_{u,1}(x_i) = \frac{1}{\pi(1 + (u - x_i)^2)}.$$

L'estimation du maximum de vraisemblance $\hat{\alpha}^*$ sera le point de maximum de $\varphi(u)$. Nous montrerons plus bas (cf. §§ 24, 25) que α^* et $\hat{\alpha}^*$ sont asymptotiquement équivalents et suivent une loi asymptotiquement normale

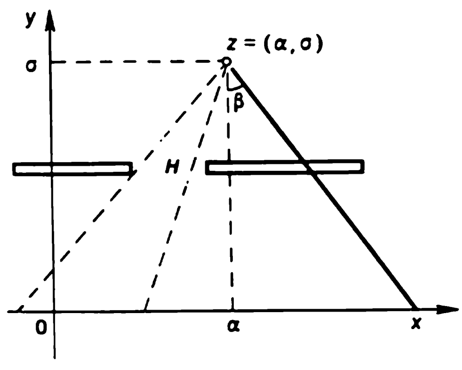


Fig. 2.

de paramètre $1/I=2$ (dans le cas considéré $I = \int (k_0')^2/k_0 dx = 4\pi^{-1}\int x^2(1+x^2)^{-3}dx = 1/2$). De ce qui précède il s'ensuit que l'erreur des estimateurs α^* et $\hat{\alpha}^*$ est de l'ordre de $1/\sqrt{n}$ pour les grands n.

Il est intéressant de noter que dans ce problème on peut obtenir une plus grande précision en plaçant entre le point $z = (\alpha, 1)$ et le détecteur un écran parallèle à l'axe des abscisses et muni d'un orifice H qui seul laissera passer les rayons γ. Les positions de l'écran et de l'orifice sont choisies par l'expérimentateur et sont donc connues.

Dans ce cas la distribution des observations sur l'écran sera discontinue et pour de petits orifices H sera proche d'une distribution $U_{a\alpha,\, a\alpha+b}$ dont on connaît les constantes a et b. Dans l'exemple 2 nous avons trouvé la forme d'un estimateur équivariant efficace α_H^* pour cette distribution. L'estimateur α_H^* est défini par les valeurs extrêmes de l'échantillon et admet une précision de l'ordre de $1/n_H$, où $n_H \leqslant n$ est le nombre d'éléments de l'échantillon associés aux rayons passant par l'orifice (n_H et n sont aléatoires et suivent la loi de Poisson). Vu que n_H est en moyenne proportionnel à n, pour les n assez grands on obtient $1/n_H \ll 1/\sqrt{n}$.

§ 19*. Problème général d'estimation équivariante

Considérons le groupe G des applications mesurables g de l'espace $\mathscr{X}^n$ dans lui-même, douées des propriétés suivantes :

1) Toute application g applique $\mathscr{X}^n$ sur $\mathscr{X}^n$, c'est-à-dire que pour tout $x_2 \in \mathscr{X}^n$ il existe un $x_1 \in \mathscr{X}^n$ tel que $x_2 = gx_1$.

2) Les applications g sont bijectives.

Toute application g doit être mesurable afin que gX soit une variable aléatoire. La propriété de groupe exprime que $g_2 g_1 \in G$ si $g_1 \in G$, $g_2 \in G$; l'application identique e et l'application réciproque g^{-1} appartiennent à G (de sorte que $g^{-1}g = e$).

DÉFINITION 1. On dit qu'une famille de distributions $\{\mathbf{P}_\theta\}$ est *invariante par le groupe d'applications* G (ou pour abréger, tout simplement *invariante*) si pour tout couple (g, θ) tel que $g \in G$ et $\theta \in \Theta$ il existe un seul $\theta_g \in \Theta$ tel que la relation $X \in \mathbf{P}_\theta$ entraîne $gX \in \mathbf{P}_{\theta_g}$.

La valeur θ_g définie de façon unique par θ et g sera désignée par $\theta_g = \bar{g}\theta$. Cette définition exprime alors que

$$\mathbf{P}_\theta(gX \in A) = \mathbf{P}_{\bar{g}\theta}(X \in A).$$

Puisque la condition (A_0) est remplie en vertu de la définition 1, l'ensemble $\bar{G}$ de toutes les applications $\bar{g}$ de Θ dans lui-même a une structure de groupe. En effet, la distribution de $g_2 g_1 X$ est donnée simultanément par les distributions $\mathbf{P}_{\overline{g_2 g_1}\theta}$ et $\mathbf{P}_{\bar{g}_2 \bar{g}_1 \theta}$. La condition (A_0) entraîne que $\overline{g_2 g_1} = \bar{g}_1 \bar{g}_2$ et que $\bar{g}_1^{-1} \in \bar{G}$ (il suffit de poser $g_2 = g_1^{-1}$). Les applications $\bar{g}$ de $\bar{G}$ sont automatiquement bijectives. Cependant, G et $\bar{G}$ peuvent ne pas être isomorphes. Supposons par exemple que $X \in \Phi_{0,\sigma^2}$, $\sigma \in]0, \infty[$. Dans ce cas, la densité $f_{0,\sigma^2}(X)$ (la fonction de vraisemblance) ne dépend que de $\sum x_i^2$. Si donc G est le groupe des rotations (des transformations orthogonales de $\mathscr{X}^n$), alors les conditions de la définition 1 seront remplies, tandis que $\bar{g} = \bar{e}$ et le groupe $\bar{G}$ ne sera composé que du seul élément $\bar{e}$: l'application identique de $\Theta =]0, \infty[$ dans lui-même.

Nous laissons au lecteur le soin de vérifier à titre d'exercice que si $\{\mathbf{P}_\theta\}$ est invariante par le groupe G, elle l'est par tout sous-groupe G_1 de G.

Dans le problème général d'estimation équivariante, la comparaison des estimateurs doit être envisagée d'un point de vue un peu plus général. Jusqu'ici nous avons apprécié l'erreur d'un estimateur par la quantité $(\theta^* - \theta)^2$. Nous admettrons maintenant que l'erreur de θ^* est mesurée par une fonction $w(\theta^*, \theta)$ et que cette fonction possède la propriété d' « homogénéité » *) :

$$w(\bar{g}\theta, \bar{g}\theta^*) = w(\theta, \theta^*) \quad \text{pour tous les } \theta. \tag{1}$$

*) Cette propriété n'est pas obligatoire en théorie de l'estimation équivariante. On peut exiger seulement l'existence d'un $\tilde{g}\theta^*$ tel que $w(\bar{g}\theta, \tilde{g}\theta^*) = w(\theta, \theta^*)$ pour tous les θ (cf. [27]).

Cette propriété caractérise précisément les fonctions $w(\theta, \theta^*) = (\theta - \theta^*)^2$ pour le paramètre de translation (translation) et $w(\theta, \theta^*) = \left(\ln \frac{\theta}{\theta^*}\right)^2$ ou $\left(\frac{\theta}{\theta^*} - 1\right)^2$ pour le paramètre d'échelle (contraction).

Nous avons vu au n° 4 du § 18 que la recherche du meilleur estimateur invariant pouvait être très sensible au choix de la mesure de l'erreur $w(\theta, \theta^*)$ de l'estimateur θ^*.

Considérons maintenant le problème d'estimation pour des familles invariantes $\{\mathbf{P}_\theta\}$. Supposons qu'on dispose d'un échantillon X au vu duquel on a construit un estimateur $\theta^* = \theta^*(X)$ du paramètre θ. Si l'on considère l'échantillon $Y = gX \in \mathbf{P}_{\bar{g}\theta}$, alors $\theta^*(Y)$ sera un estimateur de $\bar{g}\theta$. Ceci étant, il est naturel de supposer que les estimateurs $\theta^*(X)$ et $\theta^*(Y)$ sont liés entre eux comme les paramètres θ et $\bar{g}\theta$ à estimer, c'est-à-dire par l'application $\bar{g}$:

$$\theta^*(Y) = \bar{g}\theta^*(X). \tag{2}$$

En vertu de (1), l'estimateur $\theta^*(Y)$ du paramètre $\bar{g}\theta$ donne lieu à la même erreur que l'estimateur $\theta^*(X)$ du paramètre θ. Nous avons donc deux problèmes « identiques » d'estimation. Les applications gX et $\bar{g}\theta$ peuvent être interprétées comme des changements de coordonnées. La relation (2) exprime alors que l'estimateur θ^* est indépendant du système de coordonnées et vérifie

$$\theta^*(X) = \bar{g}^{-1}\theta^*(gX). \tag{3}$$

En d'autres termes, si θ^* est choisi de façon à vérifier (2), peu importe alors lequel des deux problèmes d'estimation il fut résoudre, puisque les résultats acquis sur $\bar{g}\theta$ dans le deuxième problème peuvent être étendus à θ dans le premier grâce à l'égalité (3).

DÉFINITION 2. On appelle estimateur *équivariant* *) un estimateur θ^* du paramètre θ de la famille invariante $\{\mathbf{P}_\theta\}$ vérifiant (3).

Considérons un point quelconque $\theta_0 \in \Theta$ et l'ensemble des points « équivalents » $\theta = \bar{g}\theta_0$, $\bar{g} \in \bar{G}$. Les classes des points « équivalents » ainsi définies déterminent une partition de l'espace en sous-ensembles appelés *orbites*.

THÉORÈME 1. *La valeur* $\mathbf{E}_\theta w(\theta, \theta^*)$, *où* θ^* *est un estimateur équivariant, est constante sur une orbite, c'est-à-dire que*

$$\mathbf{E}_\theta w(\theta, \theta^*) = \mathbf{E}_{\bar{g}\theta} w(\bar{g}\theta, \theta^*)$$

quels que soient $\theta \in \Theta$ *et* $\bar{g} \in \bar{G}$.

---

*) De tels estimateurs sont parfois appelés invariants. Mais ce terme est moins exact. Il vaut mieux le réserver aux estimateurs tels que $\theta^*(gX) = \theta^*(X)$ (c'est-à-dire pour le cas où $\bar{g} = \bar{e}$, $\forall g$).

DÉMONSTRATION.

$$\mathsf{E}_\theta w(\theta, \theta^*(X)) = \mathsf{E}_\theta w(\bar{g}\theta, \bar{g}\theta^*(X)) =$$
$$= \mathsf{E}_\theta w(\bar{g}\theta, \theta^*(gX)) = \mathsf{E}_{\bar{g}\theta} w(\bar{g}\theta, \theta^*(X)). \blacktriangleleft$$

Si l'orbite $\{\theta : \theta = \bar{g}\theta_0, \bar{g} \in \bar{G}\}$ est confondue avec Θ (comme dans le cas des paramètres de translation et d'échelle), alors $\mathsf{E}_\theta w(\theta, \theta^*) = \text{const}$ sur Θ. Cette égalité exprime que θ^* est minimax (comparer avec le théorème 11.2), de sorte que les meilleurs estimateurs équivariants sont souvent minimax dans la classe de tous les estimateurs (pour plus de détails cf. [27]).

Des théorèmes du § 11, il s'ensuit par exemple le

THÉORÈME 2. *Si Θ est une orbite et si un estimateur équivariant θ^* est bayésien (ou la limite d'estimateurs bayésiens θ^*_N au sens de la convergence $\mathsf{E}_\theta w(\theta, \theta^*) = \lim_{N\to\infty} \mathsf{E}_\theta w(\theta, \theta^*_N)$), alors θ^* est un estimateur minimax.*

Signalons également l'importante propriété suivante des estimateurs équivariants. Il nous sera commode de désigner par $\nu(g\,dx)/\nu(dx)$ la densité de la mesure ν_g, $\nu_g(B) = \nu(gB)$, par rapport à la mesure ν en un point $x \in \mathscr{X}^n$.

THÉORÈME 3. *Supposons qu'est réalisée la condition (A_μ), que $\mu^n(g\,dx)/\mu^n(dx)$ est finie et strictement positive pour tout $g \in G$ et pour $[\mu^n]$-presque toutes les valeurs de x. Supposons par ailleurs qu'il existe un seul estimateur du maximum de vraisemblance $\hat{\theta}^*$ pour tout X. Dans ces conditions, si la famille $\{\mathbf{P}_\theta\}$ est invariante, alors $\hat{\theta}^*$ est un estimateur équivariant.*

DÉMONSTRATION. On a

$$f_{\hat{\theta}^*}(X) = \frac{\mathbf{P}_{\hat{\theta}^*(X)}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_\theta(dx)}{\mu^n(dx)} \tag{4}$$

au point $x = X$. En admettant que $Y = gX$, on peut écrire

$$f_{\hat{\theta}^*(Y)}(Y) = \frac{\mathbf{P}_{\hat{\theta}^*(Y)}(g\,dx)}{\mu^n(g\,dx)} = \max_\theta \frac{\mathbf{P}_\theta(g\,dx)}{\mu^n(g\,dx)}.$$

Puisque la famille $\{\mathbf{P}_\theta\}$ est invariante et $\mu^n(g\,dx)/\mu^n(dx) > 0$ est finie, ceci équivaut à

$$\frac{\mathbf{P}_{\bar{g}^{-1}\hat{\theta}^*(Y)}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_{\bar{g}^{-1}\theta}(dx)}{\mu^n(dx)} = \max_\theta \frac{\mathbf{P}_\theta(dx)}{\mu^n(dx)}.$$

En comparant avec (4) et en utilisant l'unicité de $\hat{\theta}^*(X)$, on trouve que $\bar{g}^{-1}\hat{\theta}^*(gX) = \hat{\theta}^*(X)$. $\blacktriangleleft$

§ 20. Inégalité intégrale de Rao-Cramer. Critères pour qu'un estimateur soit asymptotiquement bayésien et asymptotiquement minimax

Ce paragraphe aurait pu être intitulé aussi « Inégalité pour l'erreur quadratique moyenne dans le cas bayésien ». Il relève dans sa plus grande partie de la théorie de l'estimation.

Les problèmes liés à l'approche asymptotique de comparaison des estimateurs ont été examinés antérieurement. Désormais, et essentiellement dans les §§ 23 à 29, ils seront le principal objet d'étude.

1. Estimateurs efficaces et super-efficaces. Au § 16 consacré à l'inégalité de Rao-Cramer nous avons laissé ouverte l'importante question suivante. Soient réalisées les conditions (R). Pour les estimateurs sans biais, on a alors

$$\mathbf{E}_\theta(\theta^* - \theta)^2 \geqslant \frac{1}{nI(\theta)}.$$

Le second membre de cette inégalité est appelé parfois *borne de Rao-Cramer*. Cette borne est atteinte pour les estimateurs R-efficaces. La question est de savoir si le choix d'un biais approprié est susceptible d'améliorer tant soit peu les estimateurs R-efficaces ou asymptotiquement R-efficaces. Cette question concerne l'importance de la borne de Rao-Cramer et le rôle du biais.

Le fait qu'en un certain point fixe θ_0 la valeur $\mathbf{E}_\theta(\theta^* - \theta)^2$ peut être rendue bien plus petite que la borne de Rao-Cramer a déjà été discuté. Il suffit en effet de prendre $\theta^* = \theta_0$. Mais cet estimateur sera très mauvais ailleurs.

On peut citer un autre exemple moins trivial où cette amélioration n'est pas acquise au détriment d'autres points. Supposons que $X \in \Phi_{\alpha,\,1}$, $\alpha \in \Theta = [0, \infty[$. L'estimateur $\alpha^* = \bar{x}$ est alors efficace et même R-efficace. Mais l'estimateur $\alpha^{**} = \max(0, \bar{x})$ sera visiblement meilleur dans le cas où $\Theta = [0, \infty[$, puisqu'il diminue les erreurs quadratiques moyennes en remplaçant les valeurs négatives inadmissibles par 0. L'estimateur α^{**} sera manifestement biaisé : $\mathbf{E}_\alpha \alpha^{**} > \alpha$, mais au point $\alpha = 0$, on a $I(\alpha) = 1$, $\mathbf{E}_0(\alpha^*)^2 = \frac{1}{n}$, $\mathbf{E}_0(\alpha^{**})^2 = \frac{1}{2n} < \frac{1}{nI(0)}$.

L'amélioration réalisée dans cet exemple est le résultat de la restriction du domaine des valeurs de l'estimateur α^* à l'ensemble Θ. Citons encore un exemple dû à Hodges dans lequel l'amélioration de l'estimateur α^* n'est pas la conséquence d'une restriction de Θ.

Supposons encore que $X \in \Phi_{\alpha,\,1}$, $\alpha \in \Theta =]-\infty, \infty[$. En plus de l'estimateur efficace $\alpha^* = \bar{x}$, considérons pour $\beta < 1$ l'estimateur

$$\alpha^{**} = \begin{cases} \bar{x} & \text{si} \quad |\bar{x}| \geqslant n^{-1/4}, \\ \beta\bar{x} & \text{sinon.} \end{cases}$$

Il est immédiat de voir que pour $\alpha>0$, le théorème limite central entraîne

$$\mathbf{P}_\alpha(|\bar{x}| < n^{-1/4}) \leqslant \mathbf{P}_\alpha((\bar{x} - \alpha)\sqrt{n} < n^{1/4} - \alpha\sqrt{n}) \to 0$$

pour $n \to \infty$. Cette proposition est également valable pour $\alpha<0$. Donc, pour $\alpha \neq 0$, l'estimateur α^{**} est confondu avec $\beta\,\bar{x}$ sur un ensemble dont la probabilité converge vers 1 et, par suite, d'après le théorème de continuité

$$(\alpha^{**} - \alpha)\sqrt{n} \in \Phi_{0,\,1}.$$

Si $\alpha=0$,

$$\mathbf{P}_0(|\bar{x}| < n^{-1/4}) = \mathbf{P}_0(|\bar{x}\sqrt{n}| < n^{1/4}) \to 1,$$

et l'estimateur α^{**} est confondu avec $\bar{x}$ sur un ensemble de probabilité tendant vers 1, de sorte que $(\alpha^{**}-\alpha)\sqrt{n} \in \Phi_{0,\,\beta^2}$. Donc, l'estimateur α^{**} est asymptotiquement normal pour tous les α, et $(\alpha^{**}-\alpha)\sqrt{n} \in \Phi_{0,\,\sigma^2(\alpha)}$, où

$$\sigma^2(\alpha) = \begin{cases} 1 & \text{si} \quad \alpha \neq 0, \\ \beta^2 < 1 & \text{si} \quad \alpha = 0. \end{cases}$$

Au point $\alpha=0$, le paramètre de dispersion $\sigma^2(0)$ est par conséquent strictement inférieur à la borne inférieure de Rao-Cramer qui est égale à 1.

Les estimateurs asymptotiquement normaux pour lesquels $\sigma^2(\theta) < I^{-1}(\theta)$ pour certains θ s'appellent parfois *super-efficaces*.

Mais les exemples envisagés ci-dessus bousculent peu le principe somme toute exact de la préférence des estimations efficaces. Plus exactement, Le Cam a montré que l'amélioration des estimateurs mise en évidence dans ces exemples ne pouvait être réalisée qu'en un nombre peu élevé de points.

Dans ce paragraphe on montre qu'outre la relation $\inf\limits_{\theta^*} \mathbf{E}_t(\theta^* - t)^2 = 0$, valable pour tout t, l'*intégrale* de $\mathbf{E}_t(\theta^* - t)^2$ admet une *borne inférieure strictement positive indépendante de* θ^* et est étroitement liée à une intégrale analogue de la fonction $(nI(t))^{-1}$. Plus exactement, dans le cas où $\theta \in R$, nous établirons une inégalité pour

$$\inf_{\theta^*} \int \mathbf{E}_t(\theta^* - t)^2\, q(t)\, dt, \tag{1}$$

valable pour toute fonction de poids $q(t) \geqslant 0$, $\int q(t)\, dt = 1$, dont le second membre sera indépendant de θ^* (et en particulier du biais $b(t)$ figurant dans l'inégalité de Rao-Cramer) et proche de la valeur J/n, où

$$J = \int \frac{q(t)}{I(t)}\, dt. \tag{2}$$

2. Inégalités fondamentales. Avant d'énoncer les théorèmes correspondants, on remarquera que l'intégrale de (1) peut être considérée comme

l'espérance mathématique $\mathsf{E}(\theta^* - \theta)^2$ dans le cas bayésien où θ admet une distribution *a priori* de densité $q(t)$ par rapport à la mesure de Lebesgue. Dans ce cas, $J = \mathsf{E}I^{-1}(\theta)$.

Désignons par $f(x, t) = f_t(x)q(t)$ la densité de la distribution conjointe de X et de θ. La dérivée de $f_t(x)$ par rapport à t sera désignée comme précédemment par $f_t'(x)$.

Supposons par ailleurs que $N_h \subset \Theta$ est le support d'une fonction h définie sur Θ : $N_h = \{t : h(t) \neq 0\}$, et N le support de $f(x, t)$ dans $\mathscr{X}^n \times \Theta$.

THÉORÈME 1. *Supposons que $f_t(x)$ est dérivable par rapport à t et que la fonction $\sqrt{I(t)}$ est intégrable sur tout intervalle fini. Pour toute fonction $h(t)$ dérivable à support borné (c'est-à-dire nulle en dehors d'un intervalle fini) telle que $N_h \subset N_q$, on a alors l'inégalité*

$$\mathsf{E}(\theta^* - \theta)^2 \geqslant \frac{[\mathsf{E}(h(\theta)/q(\theta))]^2}{n\mathsf{E}(I(\theta)[h(\theta)/q(\theta)]^2) + \mathsf{E}[h'(\theta)/q(\theta)]^2} = \frac{\left[\int h(t)\,dt\right]^2}{n\int I(t)h^2(t)/q(t)\,dt + \int (h'(t))^2/q(t)\,dt}. \tag{3}$$

DÉMONSTRATION. La fonction $h(t)$ étant à support borné, on a

$$\int (f_t(x)h(t))'\,dt = \int d(f_t(x)h(t)) = 0,$$
$$\int t(f_t(x)h(t))'\,dt = -\int f_t(x)h(t)\,dt.$$

Pour tout θ^* il vient donc

$$\int_{\mathscr{X}^n}\int_{\Theta} (\theta^* - t)(f_t(x)h(t))'\,dt\mu^n(dx) = \int_{\mathscr{X}^n}\int_{\Theta} f_t(x)h(t)\,dt\mu^n(dx) = \int_{\Theta} h(t)\,dt. \tag{4}$$

En vertu de la condition $N_h \subset N_q$, ces intégrales peuvent être considérées comme des intégrales sur N. Nous pouvons donc multiplier et diviser l'intégrant de (4) par $f(x, t)$. Nous obtenons alors

$$\mathsf{E}\left[(\theta^* - \theta)\frac{(f_\theta(X)h(\theta))'}{f(X, \theta)}\right] = \int_{N_q} h(t)\,dt = \mathsf{E}\frac{h(\theta)}{q(\theta)}.$$

D'où il s'ensuit en vertu de l'inégalité de Cauchy-Bouniakovski

$$\mathsf{E}(\theta^* - \theta)^2 \geqslant \frac{[\mathsf{E}(h(\theta)/q(\theta))]^2}{\mathsf{E}[(f_\theta(X)h(\theta))'/(f_\theta(X)q(\theta))]^2}. \tag{5}$$

Reste à ramener cette inégalité à la forme (3). Remarquons préalablement que

$$\mathbf{E}_t\,|\,L'(X,\ t)\,| \ \leqslant\ n\sqrt{I(t)}$$

et que pour presque tous *) les t

$$\mathbf{E}_t L'(X,\ t)\ =\ 0. \tag{6}$$

La première de ces propositions résulte des relations

$$\mathbf{E}_t\,|\,L'(X,\ t)\,|\ \leqslant\ n\mathbf{E}_t\,|\,l'(x_1,\ t)\,|^{\frac{1}{2}}\ \leqslant\ n\ \{\mathbf{E}_t[l'(x_1,\ t)]^2\}^{1/2}\ =\ n\sqrt{I(t)}\,,$$

qui sont la conséquence de l'inégalité de Cauchy-Bouniakovski. Pour prouver la deuxième assertion, prenons une fonction quelconque $g(t)$ à support borné, admettant une dérivée partout continue. Alors

$$\int g(t)f_t'(X)\,dt\ =\ -\ \int g'(t)f_t(X)\,dt.$$

Par ailleurs,

$$\int|\,g(t)\,|\,\mathbf{E}_t\,|\,L'(X,\ t)\,|\,dt\ \leqslant\ n\int|\,g(t)\,|\,\sqrt{I(t)}\ dt\ <\ \infty.$$

De là il s'ensuit qu'on peut intervertir l'ordre d'intégration dans l'expression suivante :

$$\int g(t)\mathbf{E}_t L'(X,\ t)\,dt\ =\ \int_{\mathscr{X}^n}\int_{\Theta} g(t)f_t'(x)\,dt\mu^n(dx)\ =$$

$$=\ -\ \int_{\mathscr{X}^n}\int_{\Theta} g'(t)f_t(x)\,dt\mu^n(dx)\ =\ -\ \int_{\Theta} g'(t)\,dt\ =\ -\ \int_{\Theta} dg(t)\ =\ 0.$$

La réalisation de cette égalité pour toutes les g exprime la véracité de (6).

Nous pouvons transformer maintenant le second membre de (5). En omettant, pour abréger, les arguments des fonctions, on obtient

$$\mathbf{E}\left[\frac{(f_\theta(X)h(\theta))'}{f_\theta(X)q(\theta)}\right]^2\ =\ \mathbf{E}\left[L'\,\frac{h}{q}\ +\ \frac{h'}{q}\right]^2\ =\ \mathbf{E}\left[\left(\frac{h}{q}\right)^2\mathbf{E}_\theta(L')^2\right]\ +$$

$$+\ 2\mathbf{E}\left[\frac{h'h}{q^2}\ \mathbf{E}_\theta L'\right]\ +\ \mathbf{E}\left(\frac{h'}{q}\right)^2\ =\ n\mathbf{E}\left[\left(\frac{h}{q}\right)^2 I\right]\ +\ \mathbf{E}\left(\frac{h'}{q}\right)^2.$$

On s'est servi du fait qu'en vertu de (6)

$$\mathbf{E}\left[\frac{h'h}{q^2}\ \mathbf{E}_\theta L'\right]\ =\ \int_{N_q}\frac{h'h}{q}\ \mathbf{E}_t L'\,dt\ =\ 0$$

et que (cf. § 16) $\mathbf{E}_\theta(L')^2 = nI(\theta)$. ◄

---

*) Au § 16 nous avons prouvé que cette égalité avait lieu pour *tous* les t sous les conditions (R). Ici, il suffit qu'elle soit vérifiée pour *presque* tous les t.

Dans les propositions à venir nous exigerons partout que $f_t(x)$ satisfasse les conditions du théorème 1.

THÉORÈME 2. *Si la fonction* $h(t)=h_0(t)\equiv q(t)/I(t)$ *est dérivable et à support borné, alors*

$$\mathbf{E}(\theta^* - \theta)^2 \geqslant \frac{J}{n}\left(1 + \frac{H}{nJ}\right)^{-1} \geqslant \frac{J}{n} - \frac{H}{n^2}, \tag{7}$$

où

$$H = \int \left[\left(\frac{q(t)}{I(t)}\right)'\right]^2 \frac{dt}{q(t)}.$$

REMARQUE 1. Les inégalités des théorèmes 1 et 2 sont *intégrales* dans la mesure où elles se rapportent à des intégrales de $\mathbf{E}_t(\theta^* - t)^2$. De ce point de vue, les inégalités du § 16 peuvent être appelées *locales*.

DÉMONSTRATION. Cette assertion résulte directement du théorème 1, puisque le second membre de (3) se transforme en $J^2/(nJ+H)$ pour $h=q/I$. ◄

Nous voyons donc que la borne inférieure des valeurs possibles de $\mathbf{E}(\theta^* - \theta)^2$ pour les grands n ne diffère que légèrement de la borne $\dfrac{J}{n} =$ $= \displaystyle\int \frac{q(t)\,dt}{nI(t)}$ qui est égale à la valeur de $\mathbf{E}(\theta_0^* - \theta)^2$ pour l'estimateur R-efficace θ_0^*. Ceci plaide pour l'utilisation des estimateurs efficaces, car ces derniers font prendre à $\mathbf{E}(\theta^* - \theta)^2$ presque sa valeur extrémale quelle que soit la fonction q.

L'estimation (7) est inaméliorable comme le prouve l'exemple suivant.

EXEMPLE 1. Soit $X \in \Phi_{\alpha,1}$. On sait alors que $I(\theta)=1$. Supposons par ailleurs que le paramètre α est choisi au hasard et que sa densité $q(t)$, $t \in]-\infty, \infty[$, est dérivable. Le dernier membre de (7) se transforme alors en $(n+H)^{-1}$, où

$$H = \int \frac{(q')^2}{q}\,dt = \mathbf{E}[(\ln q(\alpha))\ ]^2.$$

L'estimateur bayésien α_Q^* correspondant à la distribution *a priori* $\mathbf{Q}$ de densité q et minimisant $\mathbf{E}(\alpha^* - \alpha)^2$ est égal ici à (cf. § 10)

$$\alpha_Q^* = \frac{\int tq(t)f_t(X)\,dt}{\int q(t)f_t(X)\,dt} = \frac{\int tq(t)\exp(n\bar{x}t - t^2n/2)\,dt}{\int q(t)\exp(n\bar{x}t - t^2n/2)\,dt} =$$

$$= \frac{\int tq(t)\exp(-n(\bar{x} - t)^2/2)\,dt}{\int q(t)\exp(-n(\bar{x} - t)^2/2)\,dt}. \tag{8}$$

Il est immédiat de trouver la représentation asymptotique de ce rapport et de montrer que

$$\alpha_Q^* = \bar{x} + \frac{q'(\bar{x})}{nq(\bar{x})} + O\left(\frac{1}{n^2}\right), \quad \mathsf{E}(\alpha_Q^* - \alpha)^2 = \frac{1}{n} - \frac{H}{n^2} + O\left(\frac{1}{n^3}\right).$$

Mais nous opterons pour une voie plus simple en supposant que $q(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. Il est évident alors que $H=1$ et le dernier membre de (7) est $1/(n+1)$. Or, nous avons établi dans l'exemple 11.1 que

$$\mathsf{E}(\alpha_Q^* - \alpha)^2 = \frac{1}{n+1}.$$

Ce qui prouve que les inégalités (7) et (3) ne peuvent pas être améliorées.

THÉORÈME 3. *Si l'intervalle* $]a-\epsilon, a+\epsilon[$ *est contenu dans* Θ, *alors pour tout estimateur* θ^* *on a*

$$\max_{t \in]a-\epsilon, a+\epsilon[} \mathsf{E}_t(\theta^* - t)^2 \geqslant \frac{1}{n \max\limits_{t \in]a-\epsilon, a+\epsilon[} I(t) + \pi^2 \epsilon^{-2}}.$$

DÉMONSTRATION. Utilisons l'inégalité

$$\max_{t \in]a-\epsilon, a+\epsilon[} \mathsf{E}_t(\theta^* - t)^2 \geqslant \int_{a-\epsilon}^{a+\epsilon} \mathsf{E}_t(\theta^* - t)^2 q(t)\, dt$$

qui est valable pour toute densité $q(t)$ nulle en dehors de $]a-\epsilon, a+\epsilon[$. La proposition annoncée résulte du théorème 1 si l'on y pose

$$h(t) = q(t) = \frac{1}{\epsilon} \cos^2 \frac{\pi(t-a)}{2\epsilon}, \quad |t-a| \leqslant \epsilon.$$

Alors

$$\mathsf{E}_t(\theta^* - \theta)^2 \geqslant \frac{1}{n\int I(t)q(t)\, dt + \int (q'(t))^2/q(t)\, dt},$$

où

$$\int \frac{(q'(t))^2}{q(t)}\, dt = \int_{-\epsilon}^{\epsilon} \frac{\left(\frac{\pi}{2\epsilon^2}\, 2 \cos \frac{\pi t}{2\epsilon} \sin \frac{\pi t}{2\epsilon}\right)^2 \epsilon}{\cos^2 \frac{\pi t}{2\epsilon}}\, dt = \frac{1}{\epsilon^2} \int_{-1}^{1} \pi^2 \sin^2 \frac{\pi t}{2}\, dt = \frac{\pi^2}{\epsilon^2}. \blacktriangleleft$$

Signalons que dans la classe des densités $q(t)$ dérivables la fonctionnelle $\int_{-1}^{1} (q'(t))^2/q(t)\,dt$ atteint son minimum pour $q(t)=\cos^2(\pi t/2)$.

Le théorème 3 nous dit en particulier que l'intervalle des valeurs θ pour lesquelles l'estimateur θ^* est super-efficace ne peut être de longueur supérieure à $O(1/\sqrt{n})$.

3. Inégalités dans le cas où la fonction $q(\theta)/I(\theta)$ n'est pas dérivable. Si la fonction $h_0 = q/I$ ne satisfait pas les conditions du théorème 1, on a l'assertion utile suivante qui permet d'estimer le comportement asymptotique de $\mathbf{E}(\theta^* - \theta)^2$ dans le cas général.

THÉORÈME 4. *Supposons qu'une suite de fonctions $h_\epsilon(t)$ dépendant d'un paramètre $\epsilon>0$ est telle que chaque fonction h_ϵ vérifie les conditions du théorème 1 et*

1) $h_\epsilon(t) \leqslant h_0(t)$,

2) $H(\epsilon) = \displaystyle\int \frac{(h'_\epsilon(t))^2}{q(t)}\,dt < \infty$.

Dans ces conditions, pour tout $\epsilon>0$

$$\mathbf{E}(\theta^* - \theta)^2 \geqslant \frac{\left(\int h_\epsilon(t)dt\right)^2}{nJ + H(\epsilon)}.$$

DÉMONSTRATION. Elle résulte directement du théorème 1 si l'on y pose $h = h_\epsilon$.

Le théorème 4 admet l'important corollaire suivant.

THÉORÈME 5. *Si une fonction q est Riemann-intégrable et $J<\infty$, alors*

$$\mathbf{E}(\theta^* - \theta)^2 \geqslant \frac{J}{n}(1 + \delta_n),$$

où $\delta_n = o(1)$ lorsque $n \to \infty$.

DÉMONSTRATION. Posons $\hat{q}_\epsilon(t) = \min\limits_{|u|\leq\epsilon} q(t+u)$,

$$q_\epsilon(t) = \begin{cases} \hat{q}_\epsilon(t) & \text{si } \hat{q}_\epsilon(t) \geqslant \epsilon, \\ 0 & \text{sinon,} \end{cases}$$

$$I_\epsilon(t) = \max(\epsilon, I(t)),$$

$$h_\epsilon(t) = \frac{1}{2\epsilon}\int_{t-\epsilon}^{t+\epsilon} \frac{q_\epsilon(v)}{I_\epsilon(v)}\,dv \leqslant h_0(t).$$

Il est évident que la fonction h_ϵ est à support borné et dérivable pour tout $\epsilon > 0$.

De la Riemann-intégrabilité de $q(t)$ il s'ensuit que $q_\epsilon(t)\uparrow q(t)$ presque partout pour $\epsilon \to 0$. Pour le prouver il suffit de s'assurer que

$$\int_a^b [q(t) - q_\epsilon(t)]\,dt \downarrow 0. \tag{9}$$

La Riemann-intégrabilité de $q(t)$ entraîne la convergence

$$\sum_k q_\delta(2k\delta)2\delta \uparrow \int q(t)\,dt, \quad \sum_k q_\delta((2k+1)\delta)2\delta \uparrow \int q(t)\,dt$$

lorsque $\delta \to 0$. Donc,

$$\int_a^b q_\epsilon(t)\,dt \geqslant \sum_k q_{2\epsilon}(2k\epsilon)2\epsilon =$$
$$= \frac{1}{2}\left(\sum q_{2\epsilon}(4k\epsilon)4\epsilon + \sum q_{2\epsilon}((4k+2)\epsilon)4\epsilon\right) \to \int q(t)\,dt.$$

Ce qui prouve (9) et avec elle la convergence $q_\epsilon(t)\uparrow q(t)$.

En utilisant maintenant cette convergence, on trouve que $\dfrac{q_\epsilon(t)}{I_\epsilon(t)} \uparrow h_0(t)$,

$$\int h_\epsilon(t)\,dt = \int \frac{dt}{2\epsilon}\int_{-\epsilon}^{\epsilon} \frac{q_\epsilon(t+v)}{I_\epsilon(t+v)}\,dv =$$
$$= \frac{1}{2\epsilon}\int_{-\epsilon}^{\epsilon} dv \int \frac{q_\epsilon(t)}{I_\epsilon(t)}\,dt = \int \frac{q_\epsilon(t)}{I_\epsilon(t)}\,dt \uparrow J.$$

Par ailleurs,

$$|h_\epsilon'(t)| = \frac{1}{2\epsilon}\left|\frac{q_\epsilon(t+\epsilon)}{I_\epsilon(t+\epsilon)} - \frac{q_\epsilon(t-\epsilon)}{I_\epsilon(t-\epsilon)}\right| \leqslant \frac{q(t)}{\epsilon^2},$$
$$H(\epsilon) \leqslant \int \left(\frac{q(t)}{\epsilon^2}\right)^2 q^{-1}(t)\,dt = \frac{1}{\epsilon^4}.$$

Nous pouvons appliquer maintenant le théorème 3. En posant $\epsilon = \epsilon(n) = n^{-1/5}$, $n \to \infty$, on trouve que $\epsilon(n) \to 0$ et

$$\mathsf{E}(\theta^* - \theta)^2 \geqslant \frac{\left(\int h_\epsilon(t)\,dt\right)^2}{nJ + n^{4/5}} = \frac{J}{n}(1 + o(1)). \blacktriangleleft$$

4. Quelques corollaires. Critères de bayésienneté et de minimaximalité asymptotiques. L'une des principales conclusions que l'on puisse tirer au vu des résultats de ce paragraphe est en gros la suivante. S'il existe un estimateur asymptotiquement R-efficace, aucun autre ne fournit asymptotiquement le meilleur résultat « dans l' ensemble » (ou « en moyenne »). Nous

utiliserons ce fait ultérieurement au § 25. On exhibera des critères de bayésienneté asymptotique et de minimaximalité asymptotique résultant directement des théorèmes 2 et 5.

DÉFINITION 1. On dit qu'un estimateur θ_1^* est *asymptotiquement R-bayésien* si

$$\mathsf{E}n(\theta_1^* - \theta)^2 = J + o(1) \tag{10}$$

lorsque $n \to \infty$.

Ces estimateurs réalisent asymptotiquement la borne inférieure des erreurs quadratiques moyennes, définie dans les théorèmes 2 et 5. On aurait pu les appeler aussi *estimateurs asymptotiquement R-efficaces « dans l'ensemble »* (ou « en moyenne »).

On rappelle (cf. § 11) qu'un estimateur θ_1^* est *asymptotiquement bayésien* (par rapport à une distribution Q) si pour tout autre estimateur θ^*, on a

$$\lim_{n \to \infty} \sup\, [\mathsf{E}n(\theta_1^* - \theta)^2 - \mathsf{E}n(\theta^* - \theta)^2] \leqslant 0. \tag{11}$$

COROLLAIRE 1. *Si les conditions du théorème* 1 *sont remplies et si* $q(t)$ *est Riemann-intégrable, tout estimateur asymptotiquement R-bayésien est asymptotiquement bayésien.*

DÉMONSTRATION. Soit θ_1^* un estimateur asymptotiquement R-bayésien. En vertu du théorème 5, pour tout autre estimateur θ^*, on a

$$\lim_{n \to \infty} \inf\, \mathsf{E}n(\theta^* - \theta)^2 \geqslant J.$$

Ce qui combiné à (10) entraîne (11). ◄

Il est clair aussi que s'il existe un estimateur asymptotiquement R-bayésien, tout autre estimateur asymptotiquement bayésien sera asymptotiquement R-bayésien (comparer avec les remarques suivant le théorème 16.3).

Du théorème 5 il s'ensuit également le

COROLLAIRE 2. *Supposons que les conditions du théorème* 1 *sont remplies et que* $q(t)$ est *Riemann-intégrable.* Si θ_1^* *et* θ_2^* *sont des estimateurs asymptotiquement R-bayésiens, ils sont asymptotiquement équivalents au sens suivant :*

$$\mathsf{E}n(\theta_1^* - \theta_2^*)^2 \to 0, \quad (\theta_1^* - \theta_2^*)\sqrt{n} \xrightarrow[P]{} 0,$$

où la convergence en probabilité est comprise par rapport à la distribution conjointe de X *et* θ *dans* $\mathscr{X}^n \times \Theta$.

DÉMONSTRATION. Elle est entièrement identique à celles des théorèmes 8.2 et 16.4. On part de l'égalité (8.11) qui, en vertu du théorème 5, nous donne

$$\lim_{n\to\infty}\text{.sup } \mathbf{E}n(\theta_1^* - \theta_2^*)^2 \leqslant 0. \blacktriangleleft$$

Aux §§ 8 et 11 nous avons signalé que pour comparer des estimateurs, on pouvait aussi bien se servir des valeurs *moyennes* $\int q(t)\mathbf{E}_t(\theta^* - t)^2 dt$ que des valeurs *maximales*

$$\sup_{t\in\Gamma} \mathbf{E}_t(\theta^* - t)^2, \quad \Gamma \subset \Theta.$$

Pour Γ on prend soit l'ensemble Θ tout entier, soit sa partie qui contient préliminairement la valeur inconnue de θ. On rappelle qu'un estimateur $\bar{\theta}^*$ est *minimax* si pour tout autre estimateur θ^* on a

$$\sup_{t\in\Gamma} \mathbf{E}_t(\bar{\theta}^* - t)^2 \leqslant \sup_{t\in\Gamma} \mathbf{E}_t(\theta^* - t)^2.$$

Un estimateur θ_1^* est *asymptotiquement minimax* si pour tout autre estimateur θ^*, on a

$$\lim_{n\to\infty}\text{.sup.} \sup_{t\in\Gamma} \mathbf{E}_t[\sqrt{n}(\theta_1^* - t)]^2 \leqslant \lim_{n\to\infty}\text{.inf.} \sup_{t\in\Gamma} \mathbf{E}_t[\sqrt{n}(\theta^* - t)]^2.$$

COROLLAIRE 3. *Si la quantité d'information de Fisher $I(\theta)$ existe, est continue et que pour tout intervalle $\Gamma \subset \Theta$ on ait*

$$\lim_{n\to\infty}\text{.sup.} \sup_{t\in\Gamma} \mathbf{E}_t[\sqrt{n}(\theta_1^* - t)]^2 \leqslant \sup_{t\in\Gamma} I^{-1}(t), \tag{12}$$

alors θ_1^* est asymptotiquement minimax.

DÉMONSTRATION. Il suffit de montrer que pour tout estimateur θ^*, on a

$$\lim_{n\to\infty}\text{.inf.} \sup_{t\in\Gamma} \mathbf{E}_t[\sqrt{n}(\theta^* - t)]^2 \geqslant \sup_{t\in\Gamma} I^{-1}(t). \tag{13}$$

Pour toute distribution $\mathbf{Q}$ sur Γ de densité $q(t)$ dérivable par rapport à la mesure de Lebesgue, on a

$$\sup_{t\in\Gamma} \mathbf{E}_t[\sqrt{n}(\theta^* - t)]^2 \geqslant \int \mathbf{E}_t[\sqrt{n}(\theta^* - t)]^2 q(t)\, dt.$$

D'après le théorème 2, l'intégrale du second membre est $\geqslant J - H/n$ pour tout estimateur θ^*. Le premier membre de (13) est donc supérieur à

$$J = \int I^{-1}(t) q(t)\, dt.$$

Mais q est une densité dérivable quelconque, et pour $\epsilon > 0$ on peut toujours la choisir, eu égard à la continuité de $I^{-1}(t)$, telle que

$$J \geqslant \sup_{t \in \Gamma} I^{-1}(t) - \epsilon.$$

Ce qui prouve (13), puisque ϵ est arbitraire. ◄

Fermons ce numéro par l'importante remarque suivante : on peut circonscrire la recherche des estimateurs asymptotiquement optimaux à la classe $\tilde{K}_0$ des estimateurs asymptotiquement sans biais, introduite au § 16. Ceci résulte des considérations suivantes.

Nous avons déjà noté que le second membre de l'inégalité du théorème 5 était égal à $J/n + o(1/n)$ et ne dépendait pas du biais $b(\theta)$. Par ailleurs, si l'on utilise l'inégalité de Rao-Cramer pour construire la borne inférieure de $\mathbf{E}(\theta^* - \theta)^2$, on obtient

$$\mathbf{E}(\theta^* - \theta)^2 \geqslant \min_b \int q(t) \left[\frac{(1 + b'(t))^2}{nI(t)} + b^2(t) \right] dt.$$

On démontre (comparer avec [41]) que ce minimum étendu à tous les biais $b(\theta)$ est de la même forme $J/n + o(1/n)$ (moyennant certaines conditions sur la régularité de $q(t)$ et $I(t)$) et, chose essentielle pour nous, est atteint pour un biais $b(\theta)$ tel que

$$b'(t) = o(1) \quad \text{et} \quad b(t) = o(1/\sqrt{n})$$

lorsque $n \to \infty$.

La classe des estimateurs θ^* doués de tels biais n'est autre que $\tilde{K}_0$ (cf. § 16). Si $\theta^* \notin \tilde{K}_0$, la borne $J/n + o(1/n)$ est inaccessible. Ainsi, dans l'approche asymptotique où les estimateurs asymptotiquement normaux sont comparés à l'aide des valeurs de $\mathbf{E}(\theta^* - \theta)^2$ pour $q(t)$ et $I(t)$ régulières, on peut se borner aux estimateurs de la classe $K = K_{\Phi,2} \cap \tilde{K}_0$ (la classe $K_{\Phi,2}$ a été envisagée au § 8), puisque les estimateurs étrangers à la classe $\tilde{K}_0$ sont « inadmissibles » au sens indiqué.

5. Cas vectoriel. Si $\theta \in R^k$, on peut établir des théorèmes analogues aux précédents et tirer les mêmes conclusions que pour le cas scalaire.

En particulier, le théorème 5, l'un des plus importants de ce paragraphe, devient

$$d^2 \geqslant J/n + o(1/n),$$

où $d^2 = \| d_{ij} \|$, $d_{ij} = \mathbf{E}(\theta_i^* - \theta_i)(\theta_j^* - \theta_j)$, $J = \mathbf{E} I^{-1}(\theta)$.

Les raisonnements relatifs aux estimateurs bayésiens et minimax restent aussi en vigueur si l'on mesure l'erreur d'estimation par la fonction

$$v(\theta^*) = \mathbf{E}_\theta(\theta^* - \theta) V (\theta^* - \theta)^T,$$

où V est une matrice semi-définie positive. On devrait appeler bayésiens et minimax (ou asymptotiquement bayésiens et asymptotiquement minimax) les estimateurs dont les erreurs vérifient les inégalités correspondantes pour toute matrice semi-définie positive V.

§ 21. Distances de Kullback-Leibler, de Hellinger et du χ^2 et leurs propriétés

Le contenu de ce paragraphe est essentiel pour l'établissement des principaux résultats de la théorie asymptotique de l'estimation ainsi que des résultats du chapitre 3.

1. Définitions et propriétés fondamentales des distances. Soient **P** et **G** deux distributions sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ absolument continues par rapport à une mesure μ. Posons

$$\frac{d\mathbf{P}}{d\mu} = p, \quad \frac{d\mathbf{G}}{d\mu} = g,$$

et soit $N_{\mathbf{P}}$ le support de la distribution $\mathbf{P}$: $N_{\mathbf{P}} = \{x : p(x) > 0\}$.

DÉFINITION 1. On appelle *distance de Kullback-Leibler* entre les distributions **P** et **G** la quantité

$$\varrho_1(\mathbf{P}, \mathbf{G}) = \int_{N_{\mathbf{P}}} \ln \frac{p(x)}{g(x)} \mathbf{P}(dx) = \int_{N_{\mathbf{P}}} \ln \frac{p(x)}{g(x)} p(x)\mu(dx).$$

En fait, $\varrho_1(\mathbf{P}, \mathbf{G})$ n'est pas une distance ou une métrique au sens coutumier, puisqu'elle n'est pas une fonction symétrique de **P** et de **G**. Nous verrons néanmoins que $\varrho_1(\mathbf{P}, \mathbf{G})$ caractérise pertinemment (du point de vue statistique) l'écart entre **G** et **P**.

De l'inégalité $\ln(1+v) - v \leqslant 0$ et de la représentation

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int \left[\ln \frac{g}{p} - \left(\frac{g}{p} - 1\right)\right] p\mu(dx)$$

il s'ensuit que toujours $\varrho_1(\mathbf{P}, \mathbf{G}) \geqslant 0$. Dans le lemme 6.1 nous avons établi que $\varrho_1(\mathbf{P}, \mathbf{G}) = 0$ si seulement $\mathbf{P} = \mathbf{G}$.

DÉFINITION 2. On appelle *distance du* χ^2 entre les distributions **P** et **G** la quantité

$$\varrho_2(\mathbf{P}, \mathbf{G}) = \int_{N_{\mathbf{P}} \cup N_{\mathbf{G}}} \frac{(p(x) - q(x))^2}{p(x)} \mu(dx).$$

Cette distance est justiciable de presque toutes les remarques suivant la définition 1. L'origine de l'appellation « distance du χ^2 » apparaîtra plus loin.

DÉFINITION 3. On appelle *distance de Hellinger* entre les distributions **P** et **G** la quantité

$$\varrho_3(\mathbf{P}, \mathbf{G}) = \int_{N_P \cup N_P} (\sqrt{p(x)} - \sqrt{q(x)})^2 \, \mu(dx).$$

La distance de Hellinger est une fonction symétrique de **P** et de **G**, quant à $\sqrt{\varrho_3(\mathbf{P}, \mathbf{G})}$, elle possède toutes les propriétés d'une métrique (entre les fonctions $\sqrt{p(x)}$ et $\sqrt{g(x)}$ dans l'espace métrique $L_2(\mathscr{X}, \mu)$). Il est immédiat de voir que

$$\varrho_3(\mathbf{P}, \mathbf{G}) = 2\left(1 - \int \sqrt{pg} \, \mu(d\mathscr{X})\right) \leqslant 2. \tag{1}$$

Les trois distances introduites jouent un rôle essentiel dans les divers problèmes de statistique mathématique. Nous nous en assurerons dans une certaine mesure.

Si l'on utilise ces distances pour caractériser le degré de proximité des distributions **P** et **G** lorsque le rapport p/g est proche de 1, on constate qu'elles ont toutes le même comportement asymptotique à des facteurs multiplicatifs constants près. En effet, le développement

$$\ln \frac{g}{p} = \ln\left(1 + \left(\frac{g}{p} - 1\right)\right) = \left(\frac{g}{p} - 1\right) - \frac{1}{2}\left(\frac{g}{p} - 1\right)^2 + O\left(\left|\frac{g}{p} - 1\right|^3\right)$$

nous donne

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int \ln \frac{g}{p} \cdot p\mu(dx) \approx \frac{1}{2} \int \left(\frac{g}{p} - 1\right)^2 p\mu(dx) = \frac{1}{2} \varrho_2(\mathbf{P}, \mathbf{G}),$$

$$\varrho_2(\mathbf{P}, \mathbf{G}) = \int \frac{(p-g)^2}{p} \, \mu(dx) = \int (\sqrt{p} - \sqrt{g})^2 \left(1 + \sqrt{\frac{g}{p}}\right)^2 \mu(dx) \approx 4\varrho_3(\mathbf{P}, \mathbf{G}).$$

La dernière égalité entraîne aussi que $\varrho_2(\mathbf{P}, \mathbf{G}) \geqslant \varrho_3(\mathbf{P}, \mathbf{G})$.

Par ailleurs, $\varrho_1(\mathbf{P}, \mathbf{G}) \geqslant \varrho_3(\mathbf{P}, \mathbf{G})$. En effet, puisque $\ln(1+x) \leqslant x$, il vient

$$\ln \frac{g}{p} = 2 \ln\left(1 + \left(\sqrt{\frac{g}{p}} - 1\right)\right) \leqslant 2\left(\sqrt{\frac{g}{p}} - 1\right),$$

$$\varrho_1(\mathbf{P}, \mathbf{G}) = -\int \ln \frac{g}{p} \, p\mu(dx) \geqslant -2\left(\int \sqrt{pg} \, \mu(dx) - 1\right) = \varrho_3(\mathbf{P}, \mathbf{G}).$$

Dans la suite, nous étudierons le cas paramétrique et admettrons qu'est remplie la condition (A_μ). On se penchera sur les distances ϱ_i, $i = 1, 2, 3$, entre les distributions $\mathbf{P} = \mathbf{P}_{\theta_1}$ et $\mathbf{G} = \mathbf{P}_{\theta_2}$ dans $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$, ainsi qu'entre les distributions empiriques respectives (qui seront désignées par $\mathbf{P}^n_{\theta_1}$ et $\mathbf{P}^n_{\theta_2}$) dans $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$. (Signalons que ces distances ont un sens pour toutes les distributions et ne sont aucunement liées à la nature des espaces.) Si

$N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$, on peut écrire

$$\begin{aligned}
\varrho_1(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) &= \int \ln \frac{f_{\theta_1}}{f_{\theta_2}} f_{\theta_1} \mu(dx) = \mathbf{E}_{\theta_1} \ln \frac{f_{\theta_1}(x_1)}{f_{\theta_2}(x_1)}, \\
\varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) &= \int \frac{(f_{\theta_1} - f_{\theta_2})^2}{f_{\theta_1}} \mu(dx) = \mathbf{E}_{\theta_1} \left(\frac{f_{\theta_2}(x_1)}{f_{\theta_1}(x_1)} - 1 \right)^2, \\
\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) &= \int \left(\sqrt{f_{\theta_1}} - \sqrt{f_{\theta_2}} \right)^2 \mu(dx) = \mathbf{E}_{\theta_1} \left(\sqrt{\frac{f_{\theta_2}(x_1)}{f_{\theta_1}(x_1)}} - 1 \right)^2.
\end{aligned} \tag{2}$$

Si la condition $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$ n'est pas remplie, les distances $\varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$ et $\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$ seront strictement supérieures aux espérances mathématiques respectives dans (2).

Parallèlement à (2) signalons l'importante égalité suivante qui résulte de (1) :

$$\mathbf{E}_{\theta_1} \sqrt{f_{\theta_2}(x_1)/f_{\theta_1}(x_1)} = \int \sqrt{f_{\theta_2}(x) f_{\theta_1}(x)}\, \mu(dx) = 1 - \frac{1}{2} \varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}). \tag{3}$$

La proposition suivante établit un lien entre les distances $\varrho_i(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$ et $\varrho_i(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2})$.

THÉORÈME 1.

$$\begin{aligned}
\varrho_1(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2}) &= n \varrho_1(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}), \\
1 + \varrho_2(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2}) &= (1 + \varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}))^n, \\
1 - \frac{1}{2} \varrho_3(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2}) &= \left(1 - \frac{1}{2} \varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) \right)^n.
\end{aligned} \tag{4}$$

DÉMONSTRATION. Elle est presque évidente si l'on admet pour simplifier que $N_{P_{\theta_2}} \subset N_{P_{\theta_1}}$ (le principe des raisonnements est le même, mais les calculs sont plus volumineux dans le cas général). En effet, on peut alors se servir des égalités (2). La première des relations (4) résulte directement du fait que

$$\ln \frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} = \sum_{i=1}^{n} \ln \frac{f_{\theta_1}(x_i)}{f_{\theta_2}(x_i)}.$$

En vertu de (2) il vient par ailleurs

$$\begin{aligned}
1 + \varrho_2(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) &= \mathbf{E}_{\theta_1} (f_{\theta_2}(x_1)/f_{\theta_1}(x_1))^2, \\
1 - \varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})/2 &= \mathbf{E}_{\theta_1} \sqrt{f_{\theta_2}(x_1)/f_{\theta_1}(x_1)}.
\end{aligned}$$

Ces relations sont aussi valables pour les distances entre $\mathbf{P}^n_{\theta_1}$ et $\mathbf{P}^n_{\theta_2}$ (à condition de remplacer x_1 par X dans les seconds membres). Comme

$$\mathbf{E}_{\theta_1}\left(\frac{f_{\theta_2}(X)}{f_{\theta_1}(X)}\right)^\alpha = \mathbf{E}_{\theta_1}\prod_{i=1}^n \left(\frac{f_{\theta_2}(x_i)}{f_{\theta_1}(x_i)}\right)^\alpha = \left[\mathbf{E}_{\theta_1}\left(\frac{f_{\theta_2}(x_1)}{f_{\theta_1}(x_i)}\right)^\alpha\right]^n,$$

on en déduit (4) pour $\alpha=2$ et $\alpha=1/2$.

Nous laissons au lecteur le soin de prouver ce théorème dans le cas général (c'est-à-dire lorsque la condition $N_{\mathbf{P}_{\theta_2}}\subset N_{\mathbf{P}_{\theta_1}}$ n'est pas remplie). ◄

Le théorème 1 entraîne le

COROLLAIRE 1.

$$\varrho_3(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2}) \leqslant n\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}).$$

En effet, $1-\beta^n \leqslant (1-\beta)n$ pour tout $\beta \geqslant 0$. En posant $\beta=1-\frac{1}{2}\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2})$, on déduit de (4) que

$$\varrho_3(\mathbf{P}^n_{\theta_1}, \mathbf{P}^n_{\theta_2}) = 2(1-\beta^n) \leqslant 2(1-\beta)n = n\varrho_3(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}). ◄$$

2. Relation entre la distance de Hellinger et autres et la quantité d'information de Fisher. La distance de Hellinger est celle des trois distances introduites précédemment qui présentera le plus grand intérêt pour nous. Cependant, les principales assertions (les théorèmes 2 et 3) et les démonstrations seront de même nature pour ces trois distances. Aussi pour alléger l'exposé nous bornerons-nous à l'étude de la distance de Hellinger que nous désignerons dorénavant par

$$\varrho(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}) = \int\left(\sqrt{f_{\theta_1}} - \sqrt{f_{\theta_2}}\right)^2 \mu(dx).$$

Posons

$$r(\theta_1, \theta_2) = \varrho(\mathbf{P}_{\theta_1}, \mathbf{P}_{\theta_2}).$$

LEMME 1. *Si $f_\theta(x)$ est continue par rapport à θ pour $[\mu]$-presque toutes les valeurs de x et si $\theta_1 \neq \theta_2$, alors*

$$\liminf_{\substack{\theta'\to\theta_1\\ \theta''\to\theta_2}} \frac{r(\theta', \theta'')}{|\theta'-\theta''|^2} \geqslant \frac{r(\theta_1, \theta_2)}{|\theta_1-\theta_2|^2}. \tag{5}$$

Si la fonction $\sqrt{f_\theta(x)}$ est dérivable par rapport à θ pour $[\mu]$-presque toutes les valeurs de x, alors

$$\liminf_{\substack{\theta'\to\theta\\ \theta''\to\theta}} \frac{r(\theta', \theta'')}{|\theta'-\theta''|^2} \geqslant \frac{I(\theta)}{4}. \tag{6}$$

En outre,

$$\frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant \frac{1}{4} \int_0^1 I(\theta_1 + (\theta_2 - \theta_1)y)dy. \tag{7}$$

On admet de toute évidence que θ', θ'', θ_1, θ_2 et θ appartiennent à Θ.

DÉMONSTRATION. Pour vérifier (5), il suffit d'appliquer le lemme de Fatou et la continuité de $f_\theta(x)$ à la relation

$$\liminf_{\substack{\theta' \to \theta_1 \\ \theta'' \to \theta_2}} \frac{r(\theta', \theta'')}{|\theta' - \theta''|^2} \geqslant \int \liminf_{\substack{\theta' \to \theta_1 \\ \theta'' \to \theta_2}} \left(\frac{\sqrt{f_{\theta'}} - \sqrt{f_{\theta''}}}{\theta' - \theta''} \right)^2 \mu(dx).$$

D'où l'on déduit (6), puisque l'intégrant du second membre est égal à $(f'_\theta)^2/(4f_\theta)$ pour $\theta_1 = \theta_2 = \theta$.

Pour prouver (7), posons $a = \theta_2 - \theta_1$ et mettons l'accroissement $\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}}$ sous la forme

$$\frac{1}{2} \int_{\theta_1}^{\theta_2} \frac{f'_t}{\sqrt{f_t}} dt = \frac{a}{2} \int_0^1 \frac{f'_{\theta_1 + ay}}{\sqrt{f_{\theta_1 + ay}}} dy.$$

L'inégalité de Cauchy-Bouniakovski nous donne

$$(\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}})^2 = \frac{a^2}{4} \left[\int_0^1 \frac{f'_{\theta_1 + ay}}{\sqrt{f_{\theta_1 + ay}}} dy \right]^2 \leqslant \frac{a^2}{4} \int_0^1 \frac{(f'_{\theta_1 + ay})^2}{f_{\theta_1 + ay}} dy.$$

L'intégrant étant positif, nous pouvons intervertir l'ordre d'intégration dans les relations suivantes :

$$\frac{r(\theta_1, \theta_2)}{a^2} \leqslant \frac{1}{4} \int_{\mathscr{X}} \left(\int_0^1 \frac{(f'_{\theta_1 + ay})^2}{f_{\theta_1 + ay}} dy \right) \mu(dx) = \frac{1}{4} \int_0^1 I(\theta_1 + ay)dy.$$

Ce qui prouve l'inégalité (7). ◄

Posons $r(\Delta) = r(\theta, \theta + \Delta)$. Le lemme 1 entraîne aussitôt le

THÉORÈME 2. *Si la fonction $\sqrt{f_\theta(x)}$ est dérivable par rapport à θ pour $[\mu]$-presque toutes les valeurs de x et $I(\theta)$ est continue, alors il existe*

$$\lim_{\Delta \to 0} \frac{r(\Delta)}{\Delta^2} = \frac{I(\theta)}{4}. \tag{8}$$

REMARQUE 1. Cette assertion est valable aussi pour les distances ϱ_1 et ϱ_2 si l'on pose

$$r(\Delta) = \frac{1}{4} \varrho_2(\mathbf{P}_\theta, \mathbf{P}_{\theta+\Delta}), \quad r(\Delta) = \frac{1}{2} \varrho_1(\mathbf{P}_\theta, \mathbf{P}_{\theta+\Delta}).$$

La relation (6) se prouve dans ces conditions exactement comme dans le lemme 1. Quant à la démonstration de (8), elle peut impliquer des conditions subsidiaires de régularité (proches des conditions (R)) assurant la légitimité du passage à la limite sous le signe d'intégration.

Donc, les distances $\varrho_i(\mathbf{P}_\theta, \mathbf{P}_{\theta+\Delta})$, $i=1, 2, 3$, ont le même comportement asymptotique, et $I(\theta)$ caractérise la vitesse avec laquelle elles tendent vers 0 lorsque $\Delta \to 0$ $\left(\text{en effet, } \frac{1}{4} I(\theta) \text{ est la dérivée seconde de } r(v) \text{ au point } v=0\right)$.

Si l'on pose $r^{(n)}(\Delta) = \varrho(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta)$, on déduit des théorèmes 1 et 2 que

$$\lim_{\Delta \to 0} \frac{r^{(n)}(\Delta)}{\Delta^2} = \frac{nI(\theta)}{4}.$$

On a des relations identiques pour les distances ϱ_1 et ϱ_2.

3. Existence de bornes uniformes pour $r(\Delta)/\Delta^2$. L'existence de ces bornes nous permet d'obtenir dans la suite des estimateurs très utiles pour les moments du rapport de vraisemblance.

Pour simplifier l'exposé ou pour éviter d'introduire des conditions plus lourdes encore, on admettra souvent dans la suite qu'est remplie la condition

(A_c) : *l'ensemble Θ est compact.*

Du point de vue des applications, cette condition qui exprime que l'ensemble des paramètres est borné et fermé n'est généralement pas restrictive.

Nous utiliserons aussi la condition (A_0) introduite au § 6 et exprimant que $f_{\theta_1} \neq f_{\theta_2}$ pour $\theta_1 \neq \theta_2$. Dans ce cas, $r(\theta_1, \theta_2) > 0$ pour $\theta_1 \neq \theta_2$.

THÉORÈME 3. *Si les conditions (A_0) et (A_c) sont réunies et si $0 < I(\theta) \leqslant 4h < \infty$ pour tous les $\theta \in \Theta$, il existe une constante $g > 0$ telle que pour tous les θ_1, $\theta_2 \in \Theta$ l'on ait*

$$g \leqslant \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant h. \tag{9}$$

DÉMONSTRATION. La majoration résulte directement de (7). Montrons maintenant que

$$\inf_{\theta_1, \theta_2} \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \geqslant g > 0. \tag{10}$$

Si par absurde (10) n'est pas vraie, il existe une suite $(\theta_1^{(n)}, \theta_2^{(n)})$ telle que

$$\frac{r(\theta_1^{(n)}, \theta_2^{(n)})}{|\theta_1^{(n)} - \theta_2^{(n)}|^2} \to 0 \tag{11}$$

lorsque $n\to\infty$. En vertu de la condition (A_c), on peut admettre sans nuire à la généralité que $\theta_1^{(n)}\to\theta_1\in\Theta$, $\theta_2^{(n)}\to\theta_2\in\Theta$. Si $\theta_1\neq\theta_2$, alors (11) contredit (5), puisque $r(\theta_1, \theta_2)>0$ d'après la condition (A_0). Si $\theta_1=\theta_2=\theta$, alors (11) contredit (6), puisque $I(\theta)>0$. ◄

4. Cas vectoriel. Dans ce numéro on se propose d'établir des propositions identiques à celles des n°s 2 et 3 pour un paramètre vectoriel (le contenu du n° 1 n'est pas lié à la dimension de θ). Désignons par $\varphi(x, \theta)$ la fonction vectorielle de composantes

$$\varphi_i(x, \theta) = \frac{1}{\sqrt{f_\theta(x)}}\frac{\partial f_\theta(x)}{\partial\theta_i}.$$

La dérivée de la fonction $\sqrt{f_\theta(x)}$ suivant le vecteur unité $\omega=(\omega_1, \ldots, \omega_k)$ est alors égale à $((\sqrt{f_\theta(x)})', \omega)=(\text{grad}\sqrt{f_\theta(x)}, \omega)=\frac{1}{2}(\varphi(x, \theta), \omega)$. Dans ces notations la matrice de Fisher $I(\theta)$ est égale à

$$I(\theta) = \int\varphi^T(x, \theta)\varphi(x, \theta)\mu(dx).$$

Désignons par $|u|$ la norme euclidienne du vecteur $u=(u_1, \ldots, u_k)$.

Le lemme 1 admet la généralisation suivante au cas vectoriel.

LEMME 1A. *La première assertion du lemme* 1 (*cf.* (5)) *reste entièrement en vigueur pour* $k>1$.

Si la fonction $\sqrt{f_\theta(x)}$ *est dérivable par rapport à* θ *pour* $[\mu]$*-presque toutes les valeurs de* x, $\theta'\to\theta$, $\theta''=\theta'+\omega''\delta$, $\omega''\to\omega$, $|\omega''|=|\omega|=1$, $\delta\to0$, *alors*

$$\text{lim. inf}\,\frac{r(\theta', \theta'')}{|\theta'-\theta''|^2}\geqslant\frac{1}{4}\,\omega I(\theta)\omega^T. \tag{12}$$

Si en outre ω *est un vecteur unité colinéaire à* $\theta_2-\theta_1$, *de sorte que* $\theta_2=\theta_1+a\omega$, $a=|\theta_2-\theta_1|$, *alors*

$$\frac{r(\theta_1, \theta_2)}{|\theta_1-\theta_2|^2}\leqslant\frac{1}{4}\int_0^1\omega I(\theta_1+a\omega y)\omega^T dy. \tag{13}$$

DÉMONSTRATION. La première assertion du lemme 1 n'est pas liée à la dimension. La deuxième résulte du lemme de Fatou et des relations

$$\text{lim. inf}\,\frac{r(\theta', \theta'')}{|\theta'-\theta''|^2}\geqslant\int\text{lim. inf}\,\frac{(\sqrt{f_{\theta'}}-\sqrt{f_{\theta''}})^2}{|\theta'-\theta''|^2}\,\mu(dx) =$$

$$=\frac{1}{4}\int(\varphi(x, \theta), \omega)^2\,\mu(dx) = \frac{1}{4}\,\omega I(\theta)\omega^T.$$

Pour prouver (13), on remarquera que

$$\sqrt{f_{\theta_2}} - \sqrt{f_{\theta_1}} = \frac{1}{2}\int_0^a (\varphi(x, \theta_1 + y\omega), \omega)dy = \frac{a}{2}\int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega)dy;$$

$$r(\theta_1, \theta_2) = \frac{a^2}{4}\int_{\mathscr{X}}\left[\int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega)dy\right]^2 \mu(dx) \leqslant$$

$$\leqslant \frac{a^2}{4}\int_{\mathscr{X}}\int_0^1 (\varphi(x, \theta_1 + ay\omega), \omega)^2 dy\mu(dx) =$$

$$= \frac{a^2}{4}\int_0^1\int_{\mathscr{X}} (\varphi(x, \theta_1 + ay\omega), \omega)^2 \mu(dx)dy =$$

$$= \frac{a^2}{4}\int_0^1 \omega I(\theta_1 + ay\omega)\,\omega^T dy. \blacktriangleleft$$

Posons comme précédemment $r(\Delta) = r(\theta, \theta + \Delta)$. Le lemme 1A entraîne le

THÉORÈME 2A. *Si la fonction $\sqrt{f_\theta(x)}$ est dérivable pour $[\mu]$-presque toutes les valeurs de x et la matrice $I(\theta)$, continue, alors la limite*

$$\lim_{\delta \to 0} \frac{r(\delta\omega)}{\delta^2} = \frac{1}{4}\,\omega I(\theta)\omega^T$$

existe pour tout vecteur unité ω.

Comme dans le cas scalaire, le lemme 1A admet le corollaire suivant.

THÉORÈME 3A. *Si les conditions (A_0) et (A_c) sont remplies et si la matrice $I(\theta)$ est définie positive dans Θ, $4h \equiv \sup_\theta \operatorname{Tr} I(\theta) < \infty$, il existe alors une constante $g > 0$ telle que pour tous $\theta_1, \theta_2 \in \Theta$*

$$g \leqslant \frac{r(\theta_1, \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant h. \tag{14}$$

DÉMONSTRATION. Désignons par $\Lambda_1(\theta)$ et $\Lambda_k(\theta)$ respectivement la plus petite et la plus grande valeur propre de la matrice $I(\theta)$, de sorte que pour $|\omega| = 1$

$$\Lambda_1(\theta) \leqslant \omega I(\theta)\omega^T \leqslant \Lambda_k(\theta). \tag{15}$$

Par hypothèse, $\Lambda_1(\theta)>0$ partout sur Θ. Comme $(\varphi,\ \omega)^2 \leqslant |\varphi|^2 = \sum_{j=1}^{k} \varphi_j^2$, il vient

$$\int_{\mathscr{X}} (\varphi,\ \omega)^2 \mu(dx) = \omega I(\theta)\omega^T \leqslant \operatorname{Tr} I(\theta)$$

et par suite, $\Lambda_k(\theta) \leqslant \operatorname{Tr} I(\theta) \leqslant 4h$. L'inégalité (13) entraîne

$$\frac{r(\theta_1,\ \theta_2)}{|\theta_1 - \theta_2|^2} \leqslant \frac{1}{4}\int_0^1 \Lambda_k(\theta_1 + ay\omega)dy \leqslant h.$$

Prouvons maintenant la deuxième inégalité de (14). Supposons qu'elle est fausse. Comme dans le théorème 3, il existe alors une suite $(\theta_1^{(n)},\ \theta_2^{(n)})$, $\theta_1^{(n)} \to \theta_1 \in \Theta$, $\theta_2^{(n)} \to \theta_2 \in \Theta$, pour laquelle (11) est vraie. Si $\theta_1 \neq \theta_2$, ceci contredira (5). Si $\theta_1 = \theta_2 = \theta$, on peut sans nuire à la généralité admettre, eu égard à la compacité de la sphère $|\omega| = 1$, que $\theta_2^{(n)} = \theta_1^{(n)} + \delta\omega^{(n)}$, $\omega^{(n)} \to \omega$, $|\omega^{(n)}| = |\omega| = 1$. Mais dans ce cas, (11) contredira (12) et (15). ◄

5*. Relation entre les distances envisagées et les estimations. Considérons la distance de Kullback-Leibler entre une distribution $\mathbf{P}_\theta$ et une distribution $\mathbf{G}$ indépendante de θ :

$$\varrho_1(\mathbf{G},\ \mathbf{P}_\theta) = \int \ln \frac{d\mathbf{G}}{d\mu}\, \mathbf{G}(dx) - \int \ln f_\theta(x)\mathbf{G}(dx).$$

Seul le second terme

$$d(\mathbf{P}_\theta,\ \mathbf{G}) = -\int \ln f_\theta(x)\mathbf{G}(dx)$$

dépend de θ. Rappelons par ailleurs qu'au § 6 nous avons défini l'estimation du maximum de vraisemblance comme la valeur de θ qui minimise $d(\mathbf{P}_\theta,\ \mathbf{P}_n^*)$. Si la distribution de x_1 est discrète et μ est une mesure cardinale, l'expression

$$d(\mathbf{P}_n^*,\ \mathbf{P}_n^*) = -\int \ln \frac{d\mathbf{P}_n^*}{d\mu}\, \mathbf{P}_n^*(dx)$$

a un sens, $\varrho_1(\mathbf{P}_n^*,\ \mathbf{P}_\theta) = d(\mathbf{P}_\theta,\ \mathbf{P}_n^*) - d(\mathbf{P}_n^*,\ \mathbf{P}_n^*)$ et par suite on peut admettre que l'estimation par le maximum de vraisemblance minimise la distance de Kullback-Leibler $\varrho_1(\mathbf{P}_n^*,\ \mathbf{P}_\theta)$ entre $\mathbf{P}_\theta$ et $\mathbf{P}_n^*$. Dans le cas général, cette interprétation ne peut être adoptée que conventionellement.

Pour les distributions discrètes de x_1, on peut également envisager les distances $\varrho_i(\mathbf{P}_\theta,\ \mathbf{P}_n^*)$ pour $i=2$, 3 et les estimations minimisant ces distances. Pour $i=2$ par exemple, on obtient

$$\varrho_2(\mathbf{P}_\theta,\ \mathbf{P}_n^*) = \sum_i \frac{\left(\frac{\nu_i}{n} - f_\theta(a_i)\right)^2}{f_\theta(a_i)},$$

où ν_i est le nombre des éléments de l'échantillon tombant en un point a_i tel que $f_\theta(a_i) = \mathbf{P}_\theta(\{a_i\}) > 0$. Ceci est la statistique χ^2 (cf. §§ 7, 8) et c'est pour cette raison que nous avons attribué ce nom à la distance ϱ_2.

Etant donné que les distances ϱ_i possèdent des propriétés asymptotiques voisines, les estimations qui les minimisent seront, comme nous le verrons plus loin, asymptotiquement confondues.

§ 22*. Inégalité aux différences de type Rao-Cramer

Ce paragraphe se tient un peu à l'écart de l'exposé principal. Nous allons tenter de répondre au moins partiellement à la question de savoir ce qui se passe avec la borne inférieure admissible de $\mathbf{E}_\theta(\theta^* - \theta)^2$ dans le cas irrégulier, c'est-à-dire lorsque la fonction $f_\theta(x)$ n'est pas dérivable par rapport à θ ou lorsque $I(\theta) = \infty$.

Nous commencerons par un exemple indiquant que dans ces conditions le comportement des erreurs quadratiques moyennes des estimateurs (ou de leurs variances) peut être totalement différent de celui du second membre de l'inégalité de Rao-Cramer.

EXEMPLE 1. Supposons que $X \in \mathbf{U}_{0,\,\theta}$. Les conditions (R) ne sont pas remplies puisque la fonction $f_\theta(x)$ est discontinue. On sait que la statistique $S = \max \mathrm{x}_i$ est complète et exhaustive (cf. exemple 14.3) pour cette famille. Considérons l'estimateur sans biais $\theta^* = 2\mathrm{x}_1$. L'estimateur $\theta_S^* = 2\mathbf{E}_\theta(\mathrm{x}_1 \mid S)$ sera alors efficace en vertu des résultats du § 14. Calculons $\mathbf{E}_\theta(\mathrm{x}_1 \mid S)$. Puisque $\mathbf{P}_\theta(S < z) = (z/\theta)^n$, $z \in [0, \theta]$, la statistique S admet une densité égale à nz^{n-1}/θ^n sur $[0, \theta]$ et à 0 ailleurs. Pour déterminer la distribution conditionnelle $P(B \mid s) = \mathbf{P}_\theta(\mathrm{x}_1 \in B \mid S = s)$ de la quantité x_1 sachant que $S = s$, on se servira de la règle (10.2) :

$$P(dy \mid s) = \mathbf{P}_\theta(\mathrm{x}_1 \in dy \mid S = s) = \frac{\mathbf{P}_\theta(\mathrm{x}_1 \in dy,\ S \in ds)}{\mathbf{P}_\theta(S \in ds)}.$$

Le numérateur est égal à

$$\mathbf{P}_\theta(\mathrm{x}_1 \in dy,\ S \in ds) = \begin{cases} \dfrac{dy}{\theta} \cdot \dfrac{(n-1)s^{n-2}ds}{\theta^{n-1}} & \text{pour } y < s, \\[2mm] \dfrac{ds}{\theta} \cdot \dfrac{s^{n-1}}{\theta^{n-1}}, & \text{pour } y = s, \\[2mm] 0 & \text{pour } y > s. \end{cases}$$

D'où il vient que $P(dy \mid s) = \dfrac{(n-1)dy}{ns}$ pour $0 \leqslant y < s$, $P(\{s\} \mid s) = 1/n$.

Donc,

$$\mathbf{E}_\theta(x_1 \mid S) = \int_0^\theta y \frac{n-1}{nS}\, dy + \frac{S}{n} = \frac{S(n-1)}{2n} + \frac{S}{n} = \frac{n+1}{2n} S,$$

$$\theta_S^* = S\left(1 + \frac{1}{n}\right).$$

On a

$$\mathbf{V}_\theta\theta_S^* = \mathbf{E}_\theta(\theta_S^*)^2 - \theta^2 = \int_0^\theta s^2\left(1 + \frac{1}{n}\right)^2 \frac{ns^{n-1}}{\theta^n}\, ds - \theta^2 =$$

$$= \left(\frac{(n+1)^2}{n(n+2)} - 1\right)\theta^2 = \frac{\theta^2}{n(n+2)}. \qquad (1)$$

L'estimateur θ_S^* étant efficace, pour tout estimateur sans biais θ^*, on a

$$\mathbf{V}_\theta\theta^* \geqslant \frac{\theta^2}{n(n+2)}. \qquad (2)$$

Donc, pour les grands n l'erreur quadratique moyenne $\mathbf{E}_\theta(\theta_S^* - \theta)^2$ sera de l'ordre de $1/n^2$. Du point de vue de la borne inférieure de l'inégalité de Rao-Cramer qui est de l'ordre de $1/n$, cette précision est anormalement élevée *). On démontre que c'est la précision avec laquelle on détermine n'importe lequel des points de discontinuité de $f_\theta(x)$ (interdits par les conditions (R)) au vu de l'échantillon. On a vu dans l'exemple 7.4 de l'estimation de la médiane que les points où la densité $f_\theta(x)$ était infinie pouvaient être déterminés avec une plus grande précision, de sorte que, en châtiant le langage, on peut dire que plus la régularité est violée en un point, et plus la précision avec laquelle ce point est estimé au vu de l'échantillon sera grande. Si par exemple $X \in \mathbf{P}_\theta$, où $\mathbf{P}_\theta = \frac{1}{2}\mathbf{U}_{0,\theta} + \frac{1}{2}\mathbf{I}_\theta$, $\mathbf{I}_\theta$ est une distribution concentrée au point θ, alors $\mathbf{P}_\theta(S \neq \theta) = 2^{-n}$ ($S = \max x_i$), de sorte que la variance de $\theta^* - \theta$ pour $\theta^* = S$ décroîtra exponentiellement lorsque $n \to \infty$.

Peut-on dans ces conditions indiquer la borne inférieure des variances des estimateurs ? On établira plus bas une inégalité identique à celle de Rao-Cramer, qui permettra de déterminer ces bornes sous des conditions de régularité moins astreignantes que les conditions (R).

---

*) Il existe des estimateurs de θ dont la variance est de l'ordre de $1/n$. Par exemple, pour l'estimateur $\theta^{**} = 2\bar{x}$, on a $\mathbf{E}\theta^{**} = \theta$ et $\mathbf{V}\theta^{**} = \frac{4}{n}\mathbf{V}x_1 = \frac{\theta^2}{3n}$.

On admettra seulement qu'est remplie la condition (A_μ), bien que ceci ne soit pas essentiel (cf. remarque en fin du paragraphe).

Désignons par $\Delta\varphi(\theta)$ l'accroissement de la fonction $\varphi(\theta)$ sur l'intervalle $]\theta, \theta+\Delta[$, par $N^n_{\mathbf{P}_\theta}$ le support de la distribution $\mathbf{P}_\theta$ dans $\mathscr{X}^n$: $N^n_{\mathbf{P}_\theta} = \{x : f_\theta(x) \neq 0\}$, et posons $N^n = N^n_{\mathbf{P}_\theta} \cup N^n_{\mathbf{P}_{\theta+\Delta}}$.

THÉORÈME 1 (inégalité de Chapman-Robbins). *Supposons que* $\theta \in \Theta$, $\theta + \Delta \in \Theta$, $a(\theta) = \mathbf{E}_\theta \theta^*$. *Alors pour tout* $\Delta \neq 0$, *on a*

$$\mathbf{V}_\theta \theta^* \geqslant \frac{(\Delta a(\theta))^2}{\int [\Delta f_\theta(x)]^2 / f_\theta(x) \mu^n(dx)} = \frac{(\Delta a(\theta))^2}{\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta)}, \tag{3}$$

où ϱ_2 *est la distance du* χ^2. *Pour les estimateurs sans biais, il faut remplacer le numérateur par* Δ^2.

En vertu du théorème 21.1, le dénominateur de (3) est de la forme $\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta) = (1 + r_2(\Delta))^n - 1$, où

$$r_2(\Delta) = \varrho_2(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta) = \int \frac{[\Delta f_\theta(x)]^2}{f_\theta(x)} \mu(dx). \tag{4}$$

Donc, plus la distance $\varrho_2(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta)$ est grande (à Δ fixe), et plus la borne inférieure de $\mathbf{V}\theta^*$ est petite.

Si $\mathbf{P}_{\theta+\Delta}$ est absolument continue par rapport à $\mathbf{P}_\theta$, alors $N^n_{\mathbf{P}_{\theta+\Delta}} \subset N^n_{\mathbf{P}_\theta} = N^n$ et $\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta)$ peut être mise sous la forme (cf. (21.2))

$$\varrho_2(\mathbf{P}^n_{\theta+\Delta}, \mathbf{P}^n_\theta) = \mathbf{E}_\theta \left[\frac{\Delta f_\theta(X)}{f_\theta(X)} \right]^2;$$

de façon analogue $r_2(\Delta) = \mathbf{E}_\theta \left[\frac{\Delta f_\theta(x_1)}{f_\theta(x_1)} \right]^2$.

Si la distribution $\mathbf{P}_{\theta+\Delta}$ n'est pas absolument continue par rapport à $\mathbf{P}_\theta$, il existe un sous-ensemble de $N_{\mathbf{P}_{\theta+\Delta}}$ de $\mathbf{P}_{\theta+\Delta}$-mesure strictement positive sur lequel $f_\theta(x) = 0$, de sorte que l'intégrale de (4) devient infinie et l'inégalité (3), triviale. Signalons de nouveau que l'expression $\mathbf{E}_\theta[\Delta f_\theta(X)/f_\theta(X)]^2$ comprise comme une intégrale étendue à $N_{\mathbf{P}_\theta}$ est susceptible de rester finie.

DÉMONSTRATION du théorème 1. De ce qui précède il s'ensuit que sans nuire à la généralité on peut admettre que $\mathbf{P}_{\theta+\Delta}$ est absolument continue par rapport à $\mathbf{P}_\theta$, de sorte que $N^n_{\mathbf{P}_{\theta+\Delta}} \subset N^n_{\mathbf{P}_\theta} = N^n$. Comme $f_\theta(x)$ et $f_{\theta+\Delta}(x)$ sont des densités dans $\mathscr{X}^n$, il vient

$$\int \Delta f_\theta(x) \mu^n(dx) = 0.$$

D'autre part,

$$\int \theta^* \Delta f_\theta(x) \mu^n(dx) = \Delta a(\theta).$$

D'où

$$\int_{N^n} (\theta^* - a(\theta))\Delta f_\theta(x)\mu^n(dx) = \Delta a(\theta). \tag{5}$$

L'intégrant de (5) peut être représenté sur l'ensemble N^n sous la forme du produit

$$(\theta^* - a(\theta))\sqrt{f_\theta(x)} \cdot \frac{\Delta f_\theta(x)}{\sqrt{f_\theta(x)}}.$$

En appliquant ensuite l'inégalité de Cauchy-Bouniakovski, on obtient

$$(\Delta a(\theta))^2 \leqslant \int_{N^n} (\theta^* - \theta)^2 f_\theta(x)\mu^n(dx) \cdot \int_{N^n} \frac{(\Delta f_\theta(x))^2}{f_\theta(x)} \mu^n(dx). \blacktriangleleft$$

Dans la suite, conformément à la remarque faite ci-dessus, on se bornera, comme dans la démonstration du théorème 1, au cas où $\mathbf{P}_{\theta+\Delta}$ est absolument continue par rapport à $\mathbf{P}_\theta$ (autrement, l'inégalité (3) devient triviale).

COROLLAIRE 1. *Si sont remplies les conditions de régularité assurant l'existence* (cf. remarque 21.1 suivant le théorème 21.2) *de* $\lim\limits_{\Delta\to 0} r_2(\Delta)/\Delta^2 = I(\theta)$, *alors*

$$\mathbf{V}_\theta \theta^* \geqslant \frac{(a'_+(\theta))^2}{nI(\theta)}, \tag{6}$$

où $a'_+(\theta) = \limsup\limits_{\Delta\to 0} \dfrac{\Delta a(\theta)}{\Delta}$.

Pour déduire (6) à partir du théorème 1, il suffit de remarquer seulement que l'on peut choisir la suite $\Delta\to 0$ telle que $\dfrac{\Delta a(\theta)}{\Delta} \to a'_+(\theta)$. ◄

L'inégalité (6) est, de par sa forme, une sorte de généralisation de l'inégalité de Rao-Cramer (une généralisation à vrai dire fictive, puisque les conditions de régularité indiquées entraînent visiblement l'existence de $a'(\theta)$).

Il est naturel d'appeler l'inégalité (3) *inégalité aux différences* contrairement à l'inégalité (6) que l'on pourrait appeler *inégalité différentielle.*

Si donc $r_2(\Delta) \sim I(\theta)\Delta^2$ (ceci correspond au cas où f_θ est dérivable), *l'inégalité aux différences de Chapman-Robbins entraîne l'inégalité différentielle de Rao-Cramer.*

Mais si la fonction f_θ n'est pas dérivable, le comportement de $r_2(\Delta)$ sera différent lorsque $\Delta\to 0$.

Si par exemple la fonction f_θ est dérivable partout à l'exception d'un nombre fini de points de discontinuité de $\theta = \theta(x)$, dépendant de x, on aura

alors

$$r_2(\Delta) \sim c\,|\,\Delta\,|\,. \tag{7}$$

Ceci s'établit le plus facilement sur l'exemple assez typique traité au début du paragraphe.

Soit $X \in \mathbf{U}_{0,\,\theta}$. Pour que la condition de continuité absolue de $\mathbf{P}_{\theta+\Delta}$ par rapport à $\mathbf{P}_\theta$ soit remplie, on admettra dans le cas où $\mathbf{P}_\theta = \mathbf{U}_{0,\,\theta}$ que $\Delta < 0$, $|\,\Delta\,| < \theta$. Alors

$$\Delta f_\theta(x) = \begin{cases} \dfrac{1}{\theta+\Delta} - \dfrac{1}{\theta} & \text{pour } x \in [0,\, \theta + \Delta], \\ -\dfrac{1}{\theta} & \text{pour } x \in]\theta + \Delta,\, \theta], \\ 0 & \text{pour } x \notin [0,\, \theta], \end{cases}$$

$$r_2(\Delta) = \int_0^\theta \frac{(\Delta f_\theta(x))^2}{f_\theta(x)}\,dx = \int_0^{\theta+\Delta} \left[\frac{\Delta}{\theta(\theta + \Delta)}\right]^2 \theta\,dx + \\ + \int_{\theta+\Delta}^{\theta} \frac{1}{\theta^2}\,\theta\,dx = \frac{\Delta^2}{\theta(\theta + \Delta)} + \frac{|\,\Delta\,|}{\theta}\,.$$

Le fait essentiel ici est l'existence d'un intervalle de longueur comparable à Δ sur lequel $|\,\Delta f_\theta(x)\,| > c > 0$, où c est indépendant de Δ. Ceci assure l'ordre de petitesse (7) pour $r_2(\Delta)$.

En revenant à l'exemple envisagé, on remarque que pour les estimateurs sans biais du paramètre θ

$$\mathbf{V}\theta^* \geqslant \max_\Delta \frac{\Delta^2}{\left(1 + \dfrac{|\,\Delta\,|}{\theta} + \dfrac{\Delta^2}{\theta(\theta + \Delta)}\right)^n - 1}\,.$$

Quel est l'ordre de petitesse du second membre de cette inégalité lorsque $n \to \infty$? En posant $|\,\Delta\,| = y\theta/n$, on obtient

$$\mathbf{V}\theta^* \geqslant \frac{\theta^2}{n^2} \max_y \frac{y^2}{\left(1 + \dfrac{y}{n} + \dfrac{y^2}{n(n - y)}\right)^n - 1}\,.$$

Il est clair que l'expression sous le signe max est asymptotiquement équivalente à $h = \max\limits_y y^2/(e^y - 1) \approx 0{,}65$, de sorte que

$$\mathbf{V}\theta^* \geqslant \frac{\theta^2}{n^2}\,(h + o(1)).$$

Le second membre de cette inégalité est du même ordre de petitesse que celui de l'inégalité inaméliorable (2), mais le facteur multiplicatif constant de θ^2/n^2 dans (2) est « meilleur » et est égal à 1.

Parallèlement à (7) on peut avoir affaire à d'autres vitesses de convergence de $r_2(\Delta)$ vers 0 lorsque $\Delta \to 0$. On peut par exemple obtenir la relation $r_2(\Delta) \sim c\Delta^\alpha$, $\alpha < 1$, si $f_\theta(x) \to \infty$ au voisinage d'une courbe $\theta = \theta(x) \neq \text{const}$, ou encore la relation $r_2(\Delta) \sim c\Delta^\alpha$, $1 < \alpha < 2$, si f_θ est continue par rapport à θ, n'est pas dérivable mais vérifie seulement la condition de Hölder au voisinage d'une courbe $\theta = \theta(x) \neq \text{const}$. Il est immédiat de voir que l'ordre de petitesse de

$$\max_\Delta \frac{\Delta^2}{(1 + c\Delta^\alpha)^n - 1}$$

pour $\alpha < 2$ sera défini par la valeur $\Delta = (y/cn)^{1/\alpha}$, de sorte que

$$V\theta^* \geqslant \frac{1}{(cn)^{2/\alpha}} \max_y \frac{y^{2/\alpha}}{e^y - 1} (1 + o(1)).$$

Dans le cas « régulier » où $\alpha = 2$, le maximum par rapport à y est atteint au point limite $y = 0$ ($\Delta = 0$).

Signalons en conclusion de ce paragraphe que les estimateurs de $V\theta^*$ peuvent être obtenus de façon analogue pour des distributions $\mathbf{P}_\theta$ et $\mathbf{P}_{\theta+\Delta}$ non absolument continues l'une par rapport à l'autre. A cet effet, dans (5) il faut multiplier et diviser l'intégrant par $\sqrt{f_\theta(x) + f_{\theta+\Delta}(x)}$ et non par $\sqrt{f_\theta(x)}$. La condition (A_μ) n'est pas non plus aussi essentielle, puisque les mesures $\mathbf{P}_\theta$ et $\mathbf{P}_{\theta+\Delta}$ sont toujours absolument continues par rapport à $\frac{1}{2}(\mathbf{P}_\theta + \mathbf{P}_{\theta+\Delta})$.

§ 23. Inégalités auxiliaires pour le rapport de vraisemblance. Convergence des estimateurs du maximum de vraisemblance

Aux §§ 12 à 16 nous avons examiné des problèmes liés à l'existence et à la détermination sous une forme explicite d'estimateurs efficaces et R-efficaces. Nous avons vu qu'ils n'existaient pas toujours et qu'il n'était possible de les trouver que lorsque la fonction de vraisemblance était d'une forme spéciale ou lorsque l'on connaissait la forme explicite d'une statistique exhaustive complète (la première de ces conditions entraîne souvent la seconde (cf. § 15)).

Nous allons passer maintenant à la construction des estimateurs *asymptotiquement* optimaux. Les conditions d'existence seront ici bien moins restrictives. Les résultats correspondants reposeront essentiellement sur les

propriétés asymptotiques de la fonction

$$Z(u) = \frac{f_{\theta+u}(X)}{f_\theta(X)} = \exp\{L(X, \theta + u) - L(X, \theta)\}, \tag{1}$$

où, comme précédemment, $L(X, \theta) = \sum_{i=1}^{n} l(x_i, \theta)$. Le nombre θ de (1) sera supposé en principe être fixe et représentera la vraie valeur du paramètre, c'est-à-dire une valeur telle que $X \in \mathbf{P}_\theta$. Dans ce cas, $Z(u)$ est une fonction des variables u et X et par suite sera avec la fonction de vraisemblance $f_{\theta+u}(X)$ une *fonction aléatoire* de u. La fonction $Z(u)$ sera appelée *rapport de vraisemblance*. Elle joue un rôle important en statistique mathématique. L'étude de ses propriétés constituera le principal objectif de ce paragraphe et du suivant.

On verra que $Z(u)$ est voisine de 0 à l'extérieur d'un voisinage du point $u=0$. Au voisinage de ce point la fonction $Z(u)$ se rapproche, dans un certain sens, de la fonction de Dirac, plus exactement, $Z(v/\sqrt{n})$ se rapproche asymptotiquement pour $n \to \infty$ de la densité de la loi normale.

Aux §§ 23 à 26 nous n'envisagerons que le cas *scalaire*. Le cas vectoriel sera étudié séparément au § 28.

La distance de Hellinger

$$r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta) = \int (\sqrt{f_{\theta+u}(x)} - \sqrt{f_\theta(x)})^2 \mu(dx)$$

entre les distributions $\mathbf{P}_{\theta+u}$ et $\mathbf{P}_\theta$ étudiée au § 21 jouera un rôle important dans les estimations ultérieures. On rappelle que

$$0 \leqslant r(u) = 2(1 - \int \sqrt{f_{\theta+u}(x) f_\theta(x)}\, \mu(dx)) \leqslant 2,$$

de sorte que

$$\mathbf{E}_\theta \sqrt{\frac{f_{\theta+u}(x_1)}{f_\theta(x_1)}} = \int \sqrt{f_{\theta+u}(x) f_\theta(x)}\, \mu(dx) = 1 - r(u)/2, \tag{2}$$

$$\mathbf{E}_\theta Z^{1/2}(u) = (1 - r(u)/2)^n. \tag{3}$$

S'agissant de la famille paramétrique $\{\mathbf{P}_\theta\}$ on admettra dans ce paragraphe et dans les suivants qu'*outre* (A_μ) *sont remplies les conditions* (A_0) $(f_{\theta_1}(x) \neq f_{\theta_2}(x)$ *pour* $\theta_1 \neq \theta_2)$ *et* (A_c) (Θ *est un compact*). Nous avons déjà signalé que la dernière condition était inessentielle pour les applications. Ceci est dû au fait que dans les problèmes pratiques on peut généralement indiquer, par des raisonnements *a priori*, les bornes des valeurs possibles de θ. Pour simplifier nous admettrons si besoin que Θ est convexe (en dimension un cela exprime que $\Theta = [a, b]$, $-\infty < a < b < \infty$).

On admettra en outre dans ce paragraphe que *la fonction* $\sqrt{f_\theta}$ *est dérivable pour* $[\mu]$-*presque toutes les valeurs de* x *et* que *la quantité d'informa-*

tion de Fisher

$$I(\theta) = \int \frac{(f_\theta'(x))^2}{f_\theta(x)} \mu(dx) = \mathbf{E}_\theta \left(\frac{f_\theta'(x_1)}{f_\theta(x_1)} \right)^2$$

est strictement positive et bornée dans Θ. Sous ces conditions le théorème 21.3 nous dit que pour toutes les valeurs admissibles de θ et $\theta + u$ (c'est-à-dire telles que $\theta \in \Theta$, $\theta + u \in \Theta$) la quantité $r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta)$ vérifie l'inégalité

$$\inf_{\theta, u} \frac{r(u)}{u^2} \geqslant g > 0. \tag{4}$$

1. Inégalités fondamentales. Désignons pour simplifier $p(u) = Z^{3/4}(u)$ et admettons que toutes les conditions énumérées ci-dessus sont remplies.

THÉORÈME 1.

$$\begin{gathered} \mathbf{E}_\theta Z^{1/2}(u) \leqslant e^{-ngu^2/2}, \quad \mathbf{E}_\theta p(u) \leqslant e^{-ngu^2/4}, \\ \mathbf{E}_\theta \,|\, p'(u) \,|\, \leqslant \frac{3}{4} \sqrt{nI(\theta + u)}\; e^{-u^2ng/4}. \end{gathered} \tag{5}$$

Des considérations du § 21 il découle que pour les valeurs $u = o(1)$ on peut dans ces inégalités remplacer g par des valeurs aussi proches que l'on veut de $I(\theta)$.

DÉMONSTRATION. En vertu de (3), (4), on a

$$\mathbf{E}_\theta Z^{1/2}(u) = (1 - r(u)/2)^n \leqslant \exp\{-nr(u)/2\} \leqslant \exp\{-ngu^2/2\}.$$

L'inégalité de Cauchy-Bouniakovski nous donne

$$\mathbf{E}_\theta p(u) \leqslant [\mathbf{E}_\theta Z^{1/2}(u) \cdot \mathbf{E}_\theta Z(u)]^{1/2} = [\mathbf{E}_\theta Z^{1/2}(u)]^{1/2} \leqslant e^{-u^2ng/4}.$$

En se servant encore de l'inégalité de Cauchy-Bouniakovski et de la relation

$$p'(u) = \frac{3}{4} L'(X, \theta + u) Z^{3/4}(u),$$

on trouve

$$\begin{aligned} \mathbf{E}_\theta \,|\, p'(u) \,| &= \frac{3}{4} \mathbf{E}_\theta \,|\, L'(X, \theta + u) \,|\, Z^{1/2}(u) Z^{1/4}(u) \leqslant \\ &\leqslant \frac{3}{4} [\mathbf{E}_\theta [L'(X, \theta + u)]^2 Z(u) \cdot \mathbf{E}_\theta Z^{1/2}(u)]^{1/2} \leqslant \\ &\leqslant \frac{3}{4} [\mathbf{E}_{\theta+u} [L'(X, \theta + u)]^2]^{1/2} e^{-u^2ng/4}. \blacktriangleleft \end{aligned}$$

THÉORÈME 2. *Pour tous z, $n \geqslant 1$, on a*

$$\mathbf{P}_\theta(\sup_{v > u} Z(v/\sqrt{n}) > e^z) \geqslant c e^{-3z/4} e^{-u^2 g/4},$$

où $c = 2 + 3\sqrt{\pi I_0/g}$, $I_0 = \sup_{\theta \in \Theta} I(\theta)$ *étant indépendant de* θ.

La démonstration du théorème passe par celle du

LEMME 1. *Pour tout* $x \geqslant 0$

$$\int_x^\infty e^{-v^2/2} dv \leqslant \sqrt{2\pi}\; e^{-x^2/2}.$$

DÉMONSTRATION *). La fonction caractéristique de la variable aléatoire $\xi \in \Phi_{0,1}$ est égale à $\mathbf{E} e^{it\xi} = e^{-t^2/2}$ et est définie sur le plan tout entier. En posant $t = -ix$, on obtient $\mathbf{E} e^{x\xi} = e^{x^2/2}$. D'où, grâce à l'inégalité de Tchébychev, on déduit

$$\mathbf{P}(\xi > x) = \mathbf{P}(e^{\xi x} > e^{x^2}) \leqslant e^{-x^2} \mathbf{E} e^{\xi x} = e^{-x^2/2}. \blacktriangleleft$$

DÉMONSTRATION du théorème 2. Estimons la fonction

$$H(\delta) = \mathbf{E}_\theta \sup_{v > \delta} p(v).$$

Si $v \in [\theta + \delta, b]$, alors

$$p(v - \theta) = p(\delta) + \int_\delta^{v - \theta} p'(u) du \leqslant p(\delta) + \int_\delta^{b - \theta} |p'(u)| \, du.$$

Vu que le dernier membre ne dépend pas de v, il vient

$$\sup_{u > \delta} p(u) \leqslant p(\delta) + \int_{u > \delta} |p'(u)| \, du,$$

$$H_+(\delta) \equiv \mathbf{E}_\theta \sup_{u > \delta} p(u) \leqslant \mathbf{E}_\theta p(\delta) + \int_{u > \delta} \mathbf{E}_\theta |p'(u)| \, du.$$

D'après le théorème 1 on en déduit que

$$H_+(\delta) \leqslant e^{-\delta^2 ng/4} + \frac{3}{4}\sqrt{n} \int_{u > \delta} \sqrt{I(\theta + u)}\; e^{-u^2 ng/4} du.$$

---

*) Les inégalités suivantes

$$\frac{1}{x+1} e^{-x^2/2} < \int_x^\infty e^{-v^2/2} dv < \frac{1}{x} e^{-x^2/2},$$

que le lecteur établira sans peine en comparant les dérivées des fonctions envisagées (ces fonctions prennent les mêmes valeurs en $x = \infty$) sont plus précises pour les grands x.

Le lemme 1 nous dit que

$$H_+(\delta) \leqslant e^{-ng\delta^2/4} + \frac{3}{4}\sqrt{nI_0} \int\limits_{|u|>\delta} e^{-ngu^2/4}du \leqslant$$

$$\leqslant e^{-ng\delta^2/4} + \frac{3}{4}\sqrt{2I_0/g} \int\limits_{v>\delta\sqrt{ng/2}} e^{-v^2/2}dv \leqslant e^{-\frac{ng\delta^2}{4}}\left(1 + \frac{3}{2}\sqrt{\pi I_0/g}\right)$$

Il est clair que la fonction

$$H_-(\delta) = \sup_{u \leqslant -\delta} p(u)$$

sera justiciable de la même estimation. Donc,

$$H(\delta) \leqslant H_+(\delta) + H_-(\delta) \leqslant (2 + 3\sqrt{\pi I_0/g})\, e^{-ng\delta^2/4}.$$

Reste à se servir de l'inégalité de Tchébychev

$$\mathbf{P}_\theta(\sup_{|t|>\delta} Z(t) > e^z) = \mathbf{P}_\theta(\sup_{|t|>\delta} p(t) > e^{3z/4}) \leqslant H(\delta)\, e^{-3z/4}. \blacktriangleleft$$

2. Estimations de la distribution et des moments de l'estimateur du maximum de vraisemblance. Convergence de l'estimateur du maximum de vraisemblance.

THÉORÈME 3. *Il existe des valeurs $c < \infty$, $g > 0$ telles que*

$$\mathbf{P}_\theta(\sqrt{n}(\hat{\theta}^* - \theta) \geqslant v) \leqslant ce^{-gv^2/4} \tag{6}$$

pour tous v et $n \geqslant 1$.

DÉMONSTRATION. Du théorème 2 il s'ensuit que

$$\mathbf{P}_\theta(\sup_{|t|>v/\sqrt{n}} Z(t) > 1) \leqslant ce^{-gv^2/4}.$$

Reste à appliquer la relation

$$\{\,|\hat{\theta}^* - \theta| \geqslant \delta\} = \{\sup_{|t|>\delta} Z(t) \geqslant \sup_{|t|>\delta} Z(t)\} \subset \{\sup_{|t|>\delta} Z(t) \geqslant Z(0) = 1\} \tag{7}$$

pour $\delta = v/\sqrt{n}$. ◄

COROLLAIRE 1. *Soit $u_n \to \infty$ une suite strictement croissante. Alors*

$$(\hat{\theta}^* - \theta)\sqrt{n}/u_n \underset{P}{\to} 0. \tag{8}$$

Si les u_n sont tels que pour tout $\alpha > 0$

$$\sum e^{-\alpha u_n^2} < \infty, \tag{9}$$

alors

$$(\hat{\theta}^* - \theta)\sqrt{n}/u_n \underset{\text{p.s.}}{\to} 0. \tag{10}$$

Ces relations sont visiblement plus fortes que les relations traduisant respectivement la convergence ($\hat{\theta}^* - \theta \underset{P}{\to} 0$) et la convergence forte ($\hat{\theta}^* - \theta \underset{\text{p.s.}}{\to} 0$) de l'estimateur du maximum de vraisemblance.

DÉMONSTRATION. La relation (8) résulte directement de (6) si l'on y pose $v = \delta u_n$. La relation (10) découle aussi de (6), puisque la somme des seconds membres de (6) formera une série convergente sous la condition (9). ◄

Par exemple, même une suite croissant aussi lentement que $u_n = \ln n$ vérifie la condition (9), de sorte que *)

$$(\hat{\theta}^* - \theta)\sqrt{n}/\ln n \underset{\text{p.s.}}{\to} 0.$$

COROLLAIRE 2. *Il existe une valeur $c_1 < \infty$ indépendante de n et de θ, telle que pour tout $\alpha \leqslant g/5$*

$$\mathbf{E}_\theta \exp\{\alpha(u^*)^2\} < c_1, \quad où \quad u^* = \sqrt{n}(\hat{\theta}^* - \theta). \tag{11}$$

DÉMONSTRATION. Une intégration par parties nous donne

$$\mathbf{E}e^{\alpha\xi^2} = -\int_0^\infty e^{\alpha v^2} d\mathbf{P}(|\xi| \geqslant v) = 1 + 2\alpha\int_0^\infty v e^{\alpha v^2}\mathbf{P}(|\xi| \geqslant v)dv.$$

Donc, en vertu du théorème 3

$$\mathbf{E}_\theta e^{\alpha(u^*)^2} \leqslant 1 + \frac{2g}{5}\int_0^\infty v e^{-gv^2/20} dv \equiv c_1 < \infty. \blacktriangleleft$$

§ 24. Propriétés asymptotiques du rapport de vraisemblance

Au paragraphe précédent nous avons établi une série d'inégalités pour $Z(u)$. Trouvons maintenant la distribution limite de ces fonctions aléatoires. Ceci est possible si les conditions (R) du § 16 sont remplies. Mais pour simplifier les raisonnements nous allons introduire des conditions supplémentaires qui ne sont pas toujours liées au fond des choses mais qui rendent les démonstrations plus brèves et plus limpides.

Nous désignerons les conditions introduites par le symbole (RR) pour

*) Il résulte de la remarque 25.2 que la relation (10) est encore valable pour des suites croissant bien plus lentement.

spécifier que nous avons affaire à des conditions de régularité qui renforcent les conditions (R).

CONDITIONS (RR) :

1) *Les conditions* (A_0), (A_c), (R) *sont remplies.*

2) *La fonction* $l(x, \theta)$ *est deux fois continûment dérivable par rapport à* θ *pour* $[\mu]$*-presque toutes les valeurs de* x. *La fonction* $|l''(x, t)|$ *est majorée par la fonction* $l(x)$ *indépendante de* t : $|l''(x, t)| < l(x)$, *pour laquelle l'intégrale*

$$\mathbf{E}_t l(x_1) = \int l(x) f_t(x) \mu(dx)$$

converge uniformément par rapport à $t \in \Theta$ *).

Par convergence uniforme de l'intégrale, on entend la convergence **)

$$\sup_\theta \int_{x:\, l(x) > N} l(x) f_\theta(x) \mu(dx) \to 0$$

pour $N \to \infty$.

Dans la suite nous aurons besoin des deux propriétés suivantes qui résultent de (RR) :

1) *La légitimité de la double dérivation par rapport au paramètre sous le signe d'intégration dans l'égalité*

$$\int f_\theta(x) \mu(dx) = 1,$$

qui exprime que

$$\int f'_\theta(x) \mu(dx) = 0, \quad \int f''_\theta(x) \mu(dx) = 0. \tag{1}$$

2) *La convergence uniforme de l'intégrale*

$$I(\theta) = \int (l'(x, \theta))^2 f_\theta(x) \mu(dx).$$

(Cette propriété résulte de (R) et sera utilisée dans le § 29.)

---

*) La suite de l'exposé reste entièrement en vigueur si la condition d'existence d'un majorant est affaiblie de la manière suivante : le domaine Θ peut être recouvert d'un nombre fini de domaines $\Theta_1, \ldots, \Theta_s$ de telle sorte que pour $\theta \in \Theta_j$ la fonction $|l''(x, \theta)|$ soit majorée par une fonction $l_{(j)}(x)$ indépendante de t : $|l''(x, \theta)| < l_{(j)}(x)$, pour laquelle l'intégrale

$$\mathbf{E}_\theta l_{(j)}(x_1) = \int l_{(j)}(x) f_\theta(x) \mu(dx)$$

converge uniformément par rapport à $\theta \in \Theta_j$, $j = 1, \ldots, s$.

**) Cette interprétation de la convergence uniforme est compatible avec celle du théorème 1.5.4 qui portait sur la fonction $l(x) \equiv x$. Dans le même temps ce n'est pas une convergence uniforme de $\int \varphi(x, \theta) \mu(dx)$ pour $\varphi(x, \theta) = l(x) f_\theta(x)$ lorsqu'on admet que

$$\sup_\theta \int_{x:\, |\varphi(x, \theta)| > N} \varphi(x, \theta) \mu(dx) \to 0$$

pour $N \to \infty$.

Dans le but d'alléger l'exposé nous avons reporté la démonstration de ces corollaires à l'Annexe VI. L'autre moyen de simplifier l'exposé est d'inclure les deux propriétés indiquées dans les conditions (RR) sans se soucier du fait qu'elles seront « redondantes » sous cette forme.

Puisque

$$l'(x, \theta) = \frac{f'_\theta(x)}{f_\theta(x)}, \quad l''(x, \theta) = \frac{f''_\theta(x)}{f_\theta(x)} - \left(\frac{f'_\theta(x)}{f_\theta(x)}\right)^2,$$

la relation (1) peut être mise sous la forme

$$\mathbf{E}_\theta l'(\mathrm{x}_1, \theta) = 0, \ \mathbf{E}_\theta l''(\mathrm{x}_1, \theta) = -\mathbf{E}_\theta(l'(\mathrm{x}_1, \theta))^2 = -I(\theta). \tag{2}$$

Nous nous sommes déjà servis de la première de ces égalités.

Signalons encore un corollaire des conditions (RR). Les conditions (RR) étant bien plus fortes que celles utilisées aux §§ 21 et 23, *tous les théorèmes du* § 23 *relatifs aux estimations pour la distribution de* $\sup_{|v| \geq u} Z(v/\sqrt{n})$ *et à la convergence de l'estimateur du maximum de vraisemblance sont valables.*

LEMME 1. *Si les conditions* (RR) *sont remplies, la fonction* $l''(x, \theta)$ *est continue « en moyenne » au sens suivant :*

$$\mathbf{E}_\theta \omega''_\Delta(\mathrm{x}_1) = \int \omega''_\Delta(x) f_\theta(x) \mu(dx) \to 0 \tag{3}$$

lorsque $\Delta \to 0$, *où* $\omega''_\Delta(x)$ *est le module de continuité de* $l''(x, \theta)$:

$$\omega''_\Delta(x) = \sup_{\substack{\theta \in \Theta,\ \theta + u \in \Theta \\ |u| \leq \Delta}} |l''(x, \theta + u) - l''(x, \theta)|. \tag{4}$$

DÉMONSTRATION. En vertu du théorème de la convergence dominée, la relation (3) résulte de la continuité ordinaire, puisque dans ce cas $\omega''_\Delta(x) \to 0$ pour [μ]-presque toutes les valeurs de x lorsque $\Delta \to 0$ et de plus $|\omega''_\Delta(x)| \leqslant \leqslant 2l(x)$. ◄

Posons

$$\gamma_n(\Delta, \theta) = \sup_{|v| \leq \Delta} \left| \frac{L'(X, \theta + v) - L'(X, \theta)}{nv} + I(\theta) \right|.$$

LEMME 2. *Si les conditions* (RR) *sont remplies et si* $\delta_n > 0$, $n = 1, 2, \ldots$, *est une suite convergeant vers* 0, *alors pour tout* $\theta \in \Theta$ *et* $X \in \mathbf{P}_\theta$, *on a*

$$\gamma_n(\delta_n, \theta) \underset{\text{p.s.}}{\to} 0, \quad \gamma_n(\delta_n, \theta^*) \underset{\text{p.s.}}{\to} 0.$$

Dans ces relations on peut remplacer $I(\theta)$ *par* $I(\theta^*)$ *et vice versa.*

DÉMONSTRATION. Prouvons tout d'abord la première proposition. Comme $\mathbf{E}_\theta l''(\mathrm{x}_1, \theta) = -I(\theta)$, $L''(X, \theta)/n \underset{\text{p.s.}}{\to} -I(\theta)$, il suffit de s'assurer que

$\gamma_n(\delta_n) \to 0$, où

$$\gamma_n(\Delta) = \sup_{|v| \leq \Delta} \left| \frac{L'(X, \theta+v) - L'(X, \theta)}{nv} - \frac{L''(X, \theta)}{n} \right|.$$

Mais

$$\gamma_n(\delta_n) \leqslant \sup_{|v| < \delta_n} \frac{1}{n} |L''(X, \theta+v) - L''(X, \theta)| \leqslant \frac{1}{n} \sum_{i=1}^{n} \omega''_{\delta_n}(x_i) \equiv \bar{\omega}''_{\delta_n}(X),$$

où $\omega''_\Delta(x)$ désigne le module de continuité de $l''(x, \theta)$ défini dans (4). Il est évident que pour tout $\Delta > 0$ fixe, on a

$$\bar{\omega}''_{\delta_n}(X) \leqslant \bar{\omega}''_\Delta(X),$$

pour n assez grand. Par ailleurs, la loi forte des grands nombres nous donne

$$\bar{\omega}''_\Delta(X) \underset{\text{p.s.}}{\to} \mathbf{E}_\theta \omega''_\Delta(x_1) \equiv \omega''_\Delta.$$

Le lemme 1 nous dit que $\omega''_\Delta \to 0$ lorsque $\Delta \to 0$. D'où il s'ensuit que

$$\bar{\omega}''_{\delta_n}(X) \underset{\text{p.s.}}{\to} 0, \tag{5}$$

ce qui prouve la première proposition. De (5) et de la définition de la convergence presque sûre, il s'ensuit que, outre (5),

$$\bar{\omega}''_{\delta_n + \eta_n}(X) \underset{\text{p.s.}}{\to} 0$$

pour toute suite de variables aléatoires $\eta_n \underset{\text{p.s.}}{\to} 0$. Il reste à remarquer que

$$\sup_{|v| < \delta_n} \left| \frac{L'(X, \theta^* + v) - L'(X, \theta^*)}{nv} - \frac{L''(X, \theta)}{n} \right| \leqslant \bar{\omega}''_{\delta_n + |\theta^* - \theta|}(X), \tag{6}$$

et à appliquer le corollaire 23.1. La possibilité de substituer $I(\theta^*)$ à $I(\theta)$ résulte également du corollaire 23.1 (et de la continuité de $I(\theta)$). ◄

Nous pouvons désormais formuler les propositions fondamentales relatives au comportement asymptotique du rapport de vraisemblance $Z(t)$. Posons

$$Y(u) = \ln Z(u/\sqrt{n}) = L(X, \theta + u/\sqrt{n}) - L(X, \theta)$$

et convenons de désigner par $\epsilon_n(X, \theta)$ (parfois avec d'autres indices) les suites de variables aléatoires convergeant presque sûrement vers 0 par rapport à $\mathbf{P}_\theta$.

THÉORÈME 1. *Soient remplies les conditions (RR) et soit $\delta_n > 0$ une suite quelconque convergeant vers 0. Alors pour $|u/\sqrt{n}| < \delta_n$ on a*

$$Y(u) = u\xi_n - \frac{u^2}{2} I(\theta)(1 + \epsilon_n(X, \theta, u)), \tag{7}$$

où
$$|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta) \underset{\text{p.s.}}{\to} 0, \quad \xi_n = L'(X, \theta)/\sqrt{n} \in \Phi_{0,\, I(\theta)}.$$

Le point $u^* = (\theta^* - \theta)\sqrt{n}$ de maximum de $Y(u)$ est tel que

$$u^* = \frac{\xi_n}{I(\theta)}(1 + \epsilon_n(X, \theta)), \tag{8}$$

$$2Y(u^*) = 2\ln Z(\theta^* - \theta) = \frac{\xi_n^2}{I(\theta)}(1 + \epsilon_n(X, \theta)) \in \mathbf{H}_1. \tag{9}$$

Outre (7), *on a la représentation*

$$Y(u) = Y(u^*) - \frac{(u - u^*)^2}{2} I(\theta)(1 + \epsilon_n(X, \theta, u)), \tag{10}$$

$$|\epsilon_n(X, \theta, u)| < \epsilon_n(X, \theta).$$

Dans toutes ces propositions on peut substituer $I(\theta^*)$ à $I(\theta)$.

De même que dans le lemme 2 on admet dans ce théorème que $\theta + u\sqrt{n} \in \Theta$. Cette relation sera automatiquement satisfaite pour les grands n si θ est un point intérieur de Θ.

REMARQUE 1. Il est important de remarquer que les variables aléatoires ξ_n et $\epsilon_n(X, \theta)$ figurant dans (7) sont indépendantes de u. Donc, la première proposition du théorème peut être mise sous la forme

$$\sup_{|u| \leqslant \delta_n\sqrt{n}} \left| \frac{Y(u) - u\xi_n + \frac{u^2}{2} I(\theta)}{u^2} \right| \underset{\text{p.s.}}{\to} 0.$$

Si δ_n est tel que
$$\sum e^{-ng\delta_n^2/4} < \infty, \tag{11}$$

le théorème 23.2 nous dit que

$$\sup_{|u| > \delta_n\sqrt{n}} Y(u) \underset{\text{p.s.}}{\to} -\infty$$

pour $|u| > \delta_n\sqrt{n}$.

DÉMONSTRATION du théorème 1. Du lemme 2 on déduit pour $|v| \leqslant \delta_n$

$$L'(X, \theta + v) = L'(X, \theta) - nvI(\theta)(1 + \epsilon_n(X, \theta, v)),$$
$$|\epsilon_n(X, \theta, v)| \leqslant \epsilon_n(X, \theta).$$

En intégrant par rapport à v entre 0 et $u/\sqrt{n}$, on trouve

$$L(X, \theta + u/\sqrt{n}) - L(X, \theta) = uL'(X, \theta)/\sqrt{n} - \frac{u^2}{2} I(\theta)(1 + \epsilon_n(X, \theta, u)),$$
$$|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta). \tag{12}$$

On reconnaît visiblement un développement taylorien dans lequel $L''(X, \theta)/n$ a été remplacée par $I(\theta)$ et le reste admet une estimation uniforme. Puisque

$$\xi_n \equiv \frac{1}{\sqrt{n}} L'(X, \theta) = \frac{1}{\sqrt{n}} \sum l'(x_i, \theta)$$

est une somme de variables aléatoires indépendantes équidistribuées de moyenne 0 et de variance $I(\theta)$ (cf. (2)), il s'ensuit que $\xi \in \Phi_{0,\, I(\theta)}$ en vertu du théorème limite central. Ceci prouve la représentation (7). Revenons au lemme 2 pour établir (8). Ce lemme exprime qu'il existe un ensemble A, $\mathbf{P}_\theta(A)=1$, tel que pour $X_\infty \in A$

$$\sup_{|v|<\delta_n} \left| \frac{L'(X, \theta+v)-L'(X, \theta)}{nv} + I(\theta) \right| \to 0, \quad n\to\infty. \tag{13}$$

Par ailleurs, d'après le corollaire 23.1, il existe une suite $u_n \to \infty$, $u_n/\sqrt{n} \equiv \equiv \gamma_n \to 0$ (u_n doit vérifier (23.9)) et un ensemble B, $\mathbf{P}_\theta(B)=1$, tels que pour $X_\infty \in B$ et $n\to\infty$

$$v^* = (\theta^* - \theta) = o(\gamma_n). \tag{14}$$

Vu que la suite $\delta_n \to 0$ de la relation (13) est arbitraire, cette relation sera, en vertu de (14), vérifiée au point $v=v^*$ pour $X_\infty \in A \cap B$, $\mathbf{P}_\theta(A\cap B)=1$. En se rappelant que $L'(X, \theta+v^*)=L'(X, \theta^*)=0$, on trouve que

$$\left| I(\theta) - \frac{L'(X, \theta)}{n(\theta^* - \theta)} \right| \to 0$$

pour $X_\infty \in A \cap B$. Ceci exprime que $\xi_n - I(\theta)u^* = u^* \epsilon_n(X, \theta)$. Ce que nous voulions.

En utilisant les mêmes arguments, on peut porter $u=u^*=v^*\sqrt{n}=(\theta^* - - \theta)\sqrt{n} = \frac{\xi_n}{I(\theta)} (1+\epsilon_n(X, \theta))$ dans (12). On obtient

$$L(X, \theta^*) - L(X, \theta) = \frac{\xi_n^2}{I(\theta)} (1 + \epsilon_n(X, \theta)),$$

ce qui prouve la première partie de la relation (9). La convergence de $\xi_n^2/I(\theta)$ vers la distribution χ^2 à un degré de liberté résulte des théorèmes de continuité puisque $\xi_n/\sqrt{I(\theta)} \in \Phi_{0,\,1}$.

La relation (10) se prouve exactement comme la relation (7) si l'on se sert de la deuxième proposition du lemme 2 pour trouver une représentation pour $L(X, \theta+u\sqrt{n})-L(X, \theta^*)$. ◄

REMARQUE 2. En termes de distributions la première proposition du théorème 1 peut être formulée comme suit

$$Y(u) \in \Phi_{-u^2 I(\theta)/2,\; u^2 I(\theta)}. \tag{15}$$

Nous avons signalé plus haut que la deuxième condition (*RR*) (d'existence de $l''(x, \theta)$) n'est pas toujours essentielle pour les propositions à prouver. On s'assure à l'aide des raisonnements suivants que cette condition n'est pas essentielle à la convergence de (15). La quantité

$$Y(u) = L\left(X, \theta + \frac{u}{\sqrt{n}}\right) - L(X, \theta) = \sum_{i=1}^{n}\left[l\left(x_i, \theta + \frac{u}{\sqrt{n}}\right) - l(x_i, \theta)\right]$$

est une somme de variables aléatoires indépendantes équidistribuées. Donc, le théorème limite central appliqué à un schéma de séries (les termes dépendent de n et la vérification des conditions de Lindeberg est omise) nous dit que

$$Y(u) \in \Phi_{\alpha(u),\, \sigma^2(u)},$$

où

$$\alpha(u) = \lim_{n\to\infty} n\mathbf{E}_\theta[l(x_1, \theta + u\sqrt{n}) - l(x_1, \theta)] =$$

$$= \lim_{n\to\infty} n\mathbf{E}_\theta \ln \frac{f_{\theta+u/\sqrt{n}}(x_1)}{f_\theta(x_1)} = -u^2 \lim_{\Delta\to 0} \frac{\varrho_1(\mathbf{P}_{\theta+\Delta}, \mathbf{P}_\theta)}{\Delta^2} = -u^2 I(\theta)/2$$

(cf. théorème 21.2 et remarque 21.1). Par ailleurs,

$$\sigma^2(u) = \lim_{n\to\infty} n\mathbf{E}_\theta[l(x_1, \theta + u/\sqrt{n}) - l(x_1, \theta)]^2 =$$

$$= u^2 \lim_{\Delta\to 0} \int \left[\frac{l(x, \theta+\Delta) - l(x, \theta)}{\Delta}\right]^2 f_\theta(x)\mu(dx) =$$

$$= u^2 \int (l'(x, \theta))^2 f_\theta(x)\mu(dx) = u^2 I(\theta).$$

On aurait obtenu le même résultat en calculant $\alpha(u)$ et $\sigma^2(u)$ au moyen d'un développement en série de $l(x, \theta + u/\sqrt{n})$ limité aux deux premières dérivées. Mais nous avons vu que cela n'était pas obligatoire.

Avant de fermer ce paragraphe tirons du théorème 1 un corollaire utile concernant le comportement des intégrales du rapport de vraisemblance.

THÉORÈME 2. *Supposons que les conditions* (*RR*) *sont remplies, que la fonction* $w(t)$ *satisfait la condition*

$$|w(t)| \leq ce^{\alpha|t|^2}, \quad c < \infty, \quad \alpha = g/16$$

($g > 0$ est définie dans le § 21) *et que* $q(t)$ *est une fonction bornée continue en* $t = 0$. *Supposons par ailleurs que* Π *est une mesure sur* $(R, \mathfrak{B})$ *telle que* $\int e^{-\alpha u^2/4}\Pi(du) < \infty$. *Dans ces conditions, si* θ *est un point intérieur de* Θ *et* $X \in \mathbf{P}_\theta$, *alors*

$$J \equiv \int w(u^* - u)q(\theta + u/\sqrt{n})Z(u/\sqrt{n})\Pi(du) =$$

$$= e^{Y(u^*)}q(\theta)\left[\int w(u^* - u)e^{-\frac{1}{2}(u-u^*)^2 I(\theta)}\Pi(du) + \epsilon_n(X, \theta)\right]. \qquad (16)$$

En particulier, si Π *est la mesure de Lebesgue,* $\Pi(du)=du$*, alors*

$$J = \sqrt{\frac{2\pi}{I(\theta)}}\, e^{Y(u^*)} q(\theta)(\mathbf{E}w(\eta) + \epsilon_n(X,\theta)),$$

où

$$\epsilon_n(X,\theta) \underset{\text{p.s.}}{\to} 0, \quad \eta \in \Phi_{0,\, I^{-1}(\theta)}.$$

La proposition (16) est tout à fait naturelle, puisque le facteur $q(\theta+u/\sqrt{n})$ est « presque constant » et la fonction $Z(u/\sqrt{n})=e^{Y(u)}$ se rapproche, en vertu du théorème 1, de la densité de la loi normale à une constante multiplicative près.

DÉMONSTRATION. Dans le but d'alléger l'écriture, on se bornera au cas où la mesure Π est la mesure de Lebesgue. Le passage au cas général n'apporte aucune complication.

Estimons tout d'abord la partie de l'intégrale (16) étendue au domaine $|u|>r$. Désignons-la par $J(r)$. Comme $f_\theta(X)/f_{\theta^*}(X)\leqslant 1$, en admettant pour simplifier que $Z=Z(u^*/\sqrt{n})=e^{Y(u^*)}$, $t=\theta+u/\sqrt{n}$, on obtient

$$Z^{-1}Z\left(\frac{u}{\sqrt{n}}\right) = \frac{f_t(X)}{f_{\theta^*}(X)} \leqslant \left(\frac{f_t(X)}{f_{\theta^*}(X)}\right)^{3/4} \leqslant Z^{3/4}\left(\frac{u}{\sqrt{n}}\right).$$

Donc, l'inégalité de Cauchy-Bouniakovski, le théorème 23.1 et le corollaire 23.1 nous donnent

$$\mathbf{E}_\theta w(u^* - u)Z^{-1}Z(u/\sqrt{n}) \leqslant$$
$$\leqslant [\mathbf{E}_t w^2(\sqrt{n}(\hat\theta^* - t))\mathbf{E}_0 Z^{1/2}(u/\sqrt{n})]^{1/2} \leqslant ce^{-u^2g/4}.$$

Puisque $\max q(t)<\infty$, de là et du lemme 23.1 on déduit que

$$\mathbf{E}_\theta Z^{-1}J(r) \leqslant ce^{-gr^2/4}.$$

En appliquant l'inégalité de Tchébychev, on trouve des estimations du même ordre pour $\mathbf{P}_\theta(Z^{-1}J(r)>\delta)$. Donc, si $r=r_n\to\infty$ de telle sorte que

$$\sum e^{-r_n^2 g/4} < \infty, \tag{17}$$

alors pour $y\geqslant r_n$ on a

$$Z^{-1}J(y) \underset{\text{p.s.}}{\to} 0. \tag{18}$$

Prenons $r_n=o(\sqrt{n})$ et considérons le reste $V(y)=J-J(y)$ de l'intégrale pour $y=2r_n$. D'après le théorème 1

$$Z^{-1}V(2r) = Z^{-1}\int_{|u|<2r_n} q(\theta+u/\sqrt{n})w(u^*-u)Z(u/\sqrt{n})du =$$

$$= \int_{|u|<2r_n} (q(\theta) + \epsilon_n(u))w(u^* - u) \times$$
$$\times \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \epsilon_n(X, \theta, u))\right\} du,$$

où $|\epsilon_n(u)| < \epsilon_n \to 0$, $|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta) \underset{\text{p.s.}}{\to} 0$ pour $n \to \infty$. Donc, pour établir ce théorème, il suffit, en vertu de (18), de s'assurer de la proximité des intégrales

$$\int_{|u|<2r_n} w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \epsilon_n(X, \theta, u))\right\} du,$$

$$\sqrt{\frac{2\pi}{I(\theta)}}\, \mathsf{E}w(\eta) = \int w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\} du.$$

D'après (17) et le corollaire 23.1 il existe un ensemble A, $\mathbf{P}_\theta(A) = 1$, tel que $|u^*| \leqslant r_n$ pour $X_\infty \in A$ quel que soit $n = n(X_\infty)$ assez grand. Comme $I(\theta) \geqslant g$, $|u - u^*|^2 > u^2/2$ pour $|u| > 2r_n$, $|u^*| < r_n$, on a sur l'ensemble A (cf. lemme 23.1)

$$\int_{|u|\geq 2r_n} w(u^* - u) \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\} du < ce^{-gr_n^2} \to 0.$$

Il reste donc à estimer l'intégrale

$$\int_{|u|<2r_n} w(u^* - u) \left| \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)(1 + \epsilon_n(X, \theta, u))\right\} - \right.$$
$$\left. - \exp\left\{-\frac{1}{2}(u - u^*)^2 I(\theta)\right\}\right| du \leqslant \int w(v) \left| \exp\left\{-\frac{1}{2} v^2 I(\theta) \times \right.\right.$$
$$\left.\left. \times (1 + \epsilon_n(X, \theta, v + u^*))\right\} - \exp\left\{-\frac{1}{2} v^2 I(\theta)\right\}\right| dv.$$

Or cette intégrale converge vers 0 sur l'ensemble AB, où $B = \{X_\infty : \epsilon_n(X, \theta) \to 0\}$, $\mathbf{P}_\theta(B) = 1$. Ceci résulte de la convergence vers 0 de l'intégrant quel que soit v et du fait que cet intégrant est majoré par une fonction intégrable. ◄

§ 25. Propriétés des estimateurs du maximum de vraisemblance. Normalité asymptotique. Optimalité asymptotique

Soit $X \in \mathbf{P}_\theta$ et soit $\tilde{\theta}^*$ un estimateur du maximum de vraisemblance. Les résultats des deux précédents paragraphes nous permettent de décrire entièrement les propriétés asymptotiques de θ^* lorsque la taille n de l'échantillon

croît indéfiniment. Nous établirons en outre l'un des résultats majeurs de ce chapitre, savoir que, si les conditions (RR) sont remplies, l'estimateur du maximum de vraisemblance possède toutes les propriétés d'optimalité asymptotique étudiées plus haut, c'est-à-dire est un estimateur à la fois asymptotiquement efficace, asymptotiquement bayésien (pour toute distribution *a priori* admettant une densité) et asymptotiquement minimax.

Dans ce paragraphe on admettra tacitement la réalisation des conditions (RR).

1. Normalité asymptotique de l'estimateur du maximum de vraisemblance.

THÉORÈME 1. *L'estimateur θ^* est asymptotiquement normal et de plus la convergence*

$$u^* = (\theta^* - \theta)\sqrt{n} \in \Phi_{0,\, I^{-1}(\theta)} \tag{1}$$

est réalisée simultanément pour les moments de tout ordre, c'est-à-dire qu'outre (1) *pour tout $k>0$ on a*

$$\mathbf{E}_\theta(u^*)^k \to \mathbf{E}\eta^k, \quad \eta\in\Phi_{0,\, I^{-1}\theta}. \tag{2}$$

Bien plus, pour toute fonction continue $w(t)$ telle que $|w(t)| < e^{gt^2/6}$ (cf. (23.4)), *on a*

$$\mathbf{E}_\theta w(u^*) \to \mathbf{E}w(\eta), \quad \eta \in \Phi_{0,\, I^{-1}(\theta)}. \tag{3}$$

DÉMONSTRATION. Le théorème 24.1 affirme que

$$u^* = (\theta^* - \theta)\sqrt{n} = \frac{\xi_n}{I(\theta)}(1 + \epsilon_n(X, \theta)), \tag{4}$$

où $\epsilon_n(X, \theta)\underset{p.s.}{\to} 0$, $\xi_n = L'(X, \theta)/\sqrt{n} \in \Phi_{0,\, I(\theta)}$. Ce qui prouve (1). Les relations (2) et (3) résultent de (1) et du théorème de continuité pour les moments (cf. § 1.5), puisqu'en vertu du corollaire 23.2

$$\mathbf{E}_\theta w^{6/5}(u^*) \leqslant \mathbf{E}_\theta \exp\left\{\frac{(u^*)^2 g}{5}\right\} < c < \infty. \blacktriangleleft$$

REMARQUE 1. De (1) et (2) il s'ensuit que θ^* appartient à la classe $K_{\Phi,2}$ dans laquelle la convergence $(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\, \sigma^2(\theta)}$ a lieu en même temps que celle des moments d'ordre un et deux : $\mathbf{E}_\theta(\theta^* - \theta)^2 \to \sigma^2(\theta)$. Comme déjà signalé au § 8, dans cette classe, l'approche asymptotique de comparaison des estimateurs est confondue pratiquement avec l'approche en moyenne quadratique.

REMARQUE 2. La relation (4) permet également de décrire exactement les « écarts maximaux » $(\theta^* - \theta)\sqrt{n}$ pour $n\to\infty$. Plus exactement, on sait (cf. [17], [53]) que les sommes normalisées ξ_n de variables aléatoires indépen-

dantes équidistribuées de moyenne nulle et de variance $I(\theta)$ vérifient la loi du logarithme itéré qui dit que

$$\mathbf{P}\left(\limsup_{n\to\infty} \frac{|\xi_n|}{\sqrt{2I(\theta)\ln\ln n}} = 1\right) = 1.$$

Comme dans (4) $\limsup\limits_{n\to\infty} \epsilon_n(X, \theta)=0$ presque sûrement, il vient

$$\mathbf{P}_\theta\left(\limsup_{n\to\infty} \frac{|\hat{\theta}^* - \theta|\sqrt{nI(\theta)}}{\sqrt{2\ln\ln n}} = 1\right) = 1.$$

A titre de corollaires du théorème 2 établissons maintenant quelques propriétés de l'estimateur du maximum de vraisemblance, liées à l'optimalité asymptotique.

2. Efficacité asymptotique. Au § 16 nous avons introduit la classe $\tilde{K}_0$ des estimateurs asymptotiquement sans biais, c'est-à-dire des estimateurs θ^* dont le biais $b(\theta) = \mathbf{E}_\theta\theta^* - \theta$ est tel que

$$b(\theta) = o(1/\sqrt{n}), \quad b'(\theta) = o(1). \tag{5}$$

Au § 20 nous avons exhibé des considérations qui circonscrivaient les recherches des estimateurs asymptotiquement efficaces « dans l'ensemble » à la classe $\tilde{K}_0$.

Etablissons maintenant le fait suivant.

COROLLAIRE 1. $\theta^* \in \tilde{K}_0$.

DÉMONSTRATION. La première des relations (5) résulte de (2) pour $k=1$. Pour prouver la deuxième, on remarquera que (cf. § 16)

$$1 + b'(\theta) = \mathbf{E}_\theta\hat{\theta}^* L'(X, \theta) = \mathbf{E}_\theta(\hat{\theta}^* - \theta)L'(X, \theta) =$$

$$= \mathbf{E}_\theta((\hat{\theta}^* - \theta)\sqrt{n}\xi_n) = \mathbf{E}_\theta \frac{\xi_n^2}{I(\theta)}(1 + \epsilon_n(X, \theta)),$$

$$\epsilon_n(X, \theta) \underset{\text{p.s.}}{\to} 0.$$

Si le théorème de continuité était valable ici pour les moments, on en déduirait la relation cherchée $1+b'(\theta)\to 1$ ou, ce qui est équivalent, $b'(\theta)\to 0$. Pour établir ce théorème dans le cas envisagé, il suffit de s'assurer (cf. § 1.5) que

$$\mathbf{E}_\theta|(\hat{\theta}^* - \theta)\sqrt{n}\xi_n|^{3/2} < c < \infty, \tag{6}$$

où c est indépendant de n. En se servant de l'inégalité de Hölder

$$\mathbf{E}|\xi\eta|^r \leqslant (\mathbf{E}|\xi|^{pr})^{1/p}(\mathbf{E}|\eta|^{qr})^{1/q},$$

$$p > 0, \quad q > 0, \quad \frac{1}{p} + \frac{1}{q} = 1,$$

pour $r=3/2$, $p=4$, $q=4/3$, on trouve pour le premier membre de (6) l'estimation $(\mathbf{E}_\theta[(\tilde{\theta}^* - \theta)\sqrt{n}]^6)^{1/4}(\mathbf{E}\xi_n^2)^{3/4}$ qui en vertu de (2) nous donne l'inégalité annoncée. ◄

En raison de son importance le corollaire suivant sera énoncé sous forme de théorème.

THÉORÈME 2. *L'estimateur $\tilde{\theta}^*$ est asymptotiquement R-efficace. Il est de plus asymptotiquement efficace dans $\tilde{K}_0$.*

DÉMONSTRATION. Le fait que $\tilde{\theta}^*$ est asymptotiquement R-efficace résulte directement de la définition 16.1 et de ce que

$$\mathbf{E}_\theta(\tilde{\theta}^* - \theta)^2 = \frac{1 + o(1)}{nI(\theta)}.$$

L'efficacité asymptotique dans $\tilde{K}_0$ découle du théorème 16.3. ◄

Le théorème 2 et les remarques suivant le théorème 16.3 expriment que si les conditions (RR) sont réalisées, tout estimateur asymptotiquement efficace dans $\tilde{K}_0$ est asymptotiquement R-efficace.

A noter que la restriction à $\tilde{K}_0$ de l'ensemble des estimateurs envisagés n'est pas la seule restriction pour laquelle $\tilde{\theta}^*$ devient asymptotiquement efficace.

Indiquons une autre restriction liée cette fois-ci à la propriété pour θ d'être médiane asymptotique d'une distribution d'estimateurs asymptotiquement normaux, c'est-à-dire à la propriété

$$\mathbf{P}_\theta(\theta^* > \theta) \to 1/2 \tag{7}$$

lorsque $n \to \infty$.

Désignons par $\tilde{K}^\circ$ la classe des estimateurs θ^* pour lesquels (7) est réalisée uniformément par rapport à θ. La classe $\tilde{K}^\circ$ pourrait être appelée *classe des estimateurs asymptotiquement centrés.*

THÉORÈME 3. *L'estimateur $\tilde{\theta}^*$ est de classe $\tilde{K}^\circ$ et est un estimateur asymptotiquement efficace dans la classe $\tilde{K}^\circ$.*

Nous remettons la démonstration de ce théorème au § 3.3.

3. L'estimateur du maximum de vraisemblance est asymptotiquement bayésien. Dans ce numéro, partout où l'on admettra l'existence de la densité $q(t)$ de la distribution *a priori* $\mathbf{Q}$ par rapport à la mesure de Lebesgue sur Θ on admettra en plus tacitement que cette densité est Riemann-intégrable, de sorte que les conditions du théorème 20.5 seront remplies.

THÉORÈME 4. *L'estimateur $\tilde{\theta}^*$ est asymptotiquement R-bayésien. Si $\mathbf{Q}$ est une distribution a priori de densité $q(t)$ par rapport à la mesure de*

Lebesgue, l'estimateur $\hat{\theta}^*$ est aussi un estimateur asymptotiquement bayésien associé à la distribution **Q**.

DÉMONSTRATION. Que $\hat{\theta}^*$ soit asymptotiquement R-bayésien résulte des relations

$$\lim_{n\to\infty} \mathbf{E}[\sqrt{n}(\hat{\theta}^* - \theta)]^2 = \lim_{n\to\infty} \mathbf{E}\mathbf{E}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2 =$$

$$= \mathbf{E} \lim_{n\to\infty} \mathbf{E}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2 = \mathbf{E}I^{-1}(\theta) = J.$$

Le passage à la limite sous le signe de l'espérance mathématique est licite en vertu du théorème de la convergence dominée, puisque d'après le corollaire 23.2 la quantité $\mathbf{E}_\theta[\sqrt{n}(\hat{\theta}^* - \theta)]^2$ est uniformément bornée par une constante indépendante de n et de θ.

La bayésienneté asymptotique découle du corollaire 20.1. ◄

Les remarques suivant le corollaire 20.1 et le théorème 4 entraînent que tout estimateur asymptotiquement bayésien est asymptotiquement R-bayésien.

La proposition du théorème 4 peut être renforcée. Il s'avère que les estimateurs du maximum de vraisemblance sont « presque » confondus avec les estimateurs bayésiens pour toute densité *a priori* q.

THÉORÈME 5.

$$\mathbf{E}n(\hat{\theta}^* - \theta^*_{\mathbf{Q}})^2 \to 0, \quad (\theta^*_{\mathbf{Q}} - \hat{\theta}^*)\sqrt{n} \underset{P}{\to} 0,$$

où $\theta^*_{\mathbf{Q}}$ est un estimateur bayésien associé à la distribution **Q**, *la convergence en probabilité est comprise par rapport à la distribution conjointe de X et de θ sur $\mathscr{X}^n \times \Theta$.*

Le théorème 5 découle directement du corollaire 20.2. Ce théorème est équivalent à ce que

$$\mathbf{E}_t n(\hat{\theta}^* - \theta^*_{\mathbf{Q}})^2 \to 0$$

pour presque tous les t.

On a la proposition plus forte.

THÉORÈME 6. *Soit θ un point intérieur de Θ et soit $X \in \mathbf{P}_\theta$. Si $q(t)$ est la densité d'une distribution a priori et si cette densité est continue et strictement positive à l'intérieur de Θ, alors*

$$\sqrt{n}(\hat{\theta}^* - \theta^*_{\mathbf{Q}}) \underset{\text{p.s.}}{\to} 0.$$

DÉMONSTRATION. Elle résulte du théorème 2 du paragraphe précédent. En effet

$$\theta_Q^* - \hat{\theta}^* = \frac{\int (t-\hat{\theta}^*) q(t) f_t(X) dt}{\int q(t) f_t(X) dt}.$$

En effectuant le changement de variables $t=\theta+u/\sqrt{n}$ et en divisant le numérateur et le dénominateur de cette expression par $f_\theta(X)$, on obtient

$$\theta_Q^* - \hat{\theta}^* = \frac{\int (u-u^*) q(\theta+u/\sqrt{n}) Z(u/\sqrt{n}) du}{\sqrt{n}\int q(\theta+u/\sqrt{n}) Z(u/\sqrt{n}) du}.$$

Utilisons maintenant le théorème 24.2 pour $w(t)=t$ et $w(t)=1$. Comme $\mathbf{E}w(\eta)=\mathbf{E}\eta=0$ dans le premier cas, on obtient

$$\theta_Q^* - \hat{\theta}^* = \epsilon_n(X, \theta)/\sqrt{n}, \quad \epsilon_n(X, \theta) \underset{\text{p.s.}}{\rightarrow} 0. \blacktriangleleft$$

4. L'estimateur du maximum de vraisemblance est asymptotiquement minimax.

THÉORÈME 7. *L'estimateur du maximum de vraisemblance est un estimateur asymptotiquement minimax.*

Ce théorème découle directement du corollaire 20.3 et de la proposition suivante.

LEMME 1.

$$\lim_{n\to\infty} \sup_{\theta\in\Gamma} \mathbf{E}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} I^{-1}(\theta),$$

où Γ *est un segment quelconque situé à l'intérieur de* Θ.

Le lemme 1 résulte de la convergence uniforme en à θ de l'expression (2). L'uniformité sera prouvée au § 29 (cf. n° 29.3).

§ 26*. Calcul approché des estimateurs du maximum de vraisemblance

Nous avons vu que les estimateurs les plus intéressants étaient les estimateurs efficaces, les estimateurs asymptotiquement efficaces et en particulier les estimateurs du maximum de vraisemblance. La recherche de la valeur exacte de l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est un problème assez épineux. Ceci concerne surtout les distributions ne possédant pas de statistiques exhaustives de forme relativement simple.

Par ailleurs, la recherche d'un estimateur asymptotiquement normal θ^* n'apporte aucune complication.

Nous allons indiquer une méthode de construction d'un estimateur θ_1^* asymptotiquement équivalent à un estimateur du maximum de vraisemblance $\tilde{\theta}^*$ (donc à un estimateur asymptotiquement efficace), qui est basée sur la méthode des approximations de Newton et utilise un estimateur asymptotiquement normal θ^*. Posons

$$U(t) = t - L'(X, t) \cdot (L''(X, t))^{-1}, \quad t \in \Theta,$$

$$U_1(t) = t + L'(X, t) \cdot (nI(t))^{-1}, \quad t \in \Theta.$$

THÉORÈME 1. *Si les conditions (RR) sont remplies, $X \in \mathbf{P}_\theta$ et θ^* est un estimateur asymptotiquement normal :*

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\, \sigma^2(\theta)},$$

alors l'estimateur $\theta_1^* = U(\theta^*)$ (ou $\theta_1^* = U_1(\theta^*)$) sera asymptotiquement équivalent à $\tilde{\theta}^*$, c'est-à-dire que

$$(\theta_1^* - \tilde{\theta}^*)\sqrt{n} \underset{P_\theta}{\to} 0.$$

La démonstration de ce théorème repose sur le lemme suivant.

LEMME 1. *Soient remplies les conditions (RR), $X \in \mathbf{P}_\theta$ et $\delta_n > 0$ une suite quelconque convergeant vers 0. Si θ_n est tel que $|\theta_n - \theta| \leqslant \delta_n$, alors*

$$U(\theta_n) - \tilde{\theta}^* = (\theta_n - \tilde{\theta}^*)\epsilon_n(\theta_n, \theta, X),$$

où

$$\overline{\epsilon_n} \equiv \max_{\theta_n : |\theta_n - \theta| \leq \delta_n} |\epsilon_n(\theta_n, \theta, X| \underset{P_\theta}{\to} 0.$$

On obtiendrait la même proposition en remplaçant U par une fonction U_1.

En d'autres termes, si l'on applique la méthode des approximations successives à $\tilde{\theta}^*$ et que l'on pose $\theta_0^* = \theta_n$, $\theta_1^* = U(\theta_0^*)$ (ou $\theta_1^* = U_1(\theta_0^*)$), alors $\theta_1^* - \tilde{\theta}^* = o(\theta_0^* - \tilde{\theta}^*)$, de sorte que l'approximation θ_1^* est bien meilleure que θ_0^*.

DÉMONSTRATION. Des considérations du § 24 et de la continuité de L'', il s'ensuit (cf. par exemple lemme 24.1) que

$$L'(X, \theta_n) = (\theta_n - \tilde{\theta}^*)L''(X, \tilde{\theta}), \quad L''(X, \tilde{\theta}) = n(I(\theta) + \epsilon_n'(\theta_n, \theta, X)),$$

où $\tilde{\theta} \in [\theta_n, \tilde{\theta}^*]$, $\max_{\theta_n : |\theta_n - \theta| \leq \delta_n} |\epsilon_n'(\theta_n, \theta, X| \underset{P_\theta}{\to} 0$ pour toute suite $\delta_n \to 0$. Par ailleurs,

$$L''(X, \theta_n) = n(I(\theta) + \epsilon_n''), \quad (I(\theta) + \epsilon_n')(I(\theta) + \epsilon_n'')^{-1} = 1 + \epsilon_n,$$

où ϵ_n'' et ϵ_n jouissent de la même propriété que ϵ_n'. Donc,

$$U(\theta_n) - \hat{\theta}^* = \theta_n - \hat{\theta}^* - L'(X, \theta_n)(L''(X, \theta_n))^{-1} =$$
$$= \theta_n - \hat{\theta}^* - (\theta_n - \theta^*)(1 + \epsilon_n) = (\theta_n - \hat{\theta}^*)\,\epsilon_n.$$

La démonstration est la même pour la fonction U_1. ◄

DÉMONSTRATION du théorème 1. Choisissons un $\delta_n \to 0$ tel que $\delta_n\sqrt{n} \to \infty$ et mettons $(\theta_1^* - \hat{\theta}^*)\sqrt{n}$ sous la forme

$$(U(\theta^*) - \hat{\theta}^*)\sqrt{n} = \sqrt{n}(\theta^* - \hat{\theta}^*)\epsilon_n(\theta^*, \theta, X)I_{|\theta^* - \theta| \leq \delta_n} + r_n,$$

où $r_n \neq 0$ uniquement sur l'ensemble $B_n = \{X : |\theta^* - \theta| > \delta_n\}$ et, en vertu du lemme 1,

$$\bar{\epsilon}_n = \max_{|t - \theta| \leq \delta_n} \epsilon_n(t, \theta, X) \underset{P_\theta}{\to} 0.$$

Vu que d'autre part $\mathbf{P}_\theta(B_n) \to 0$, on en déduit que

$$|\theta_1^* - \hat{\theta}^*|\sqrt{n} \leqslant \sqrt{n}|\theta^* - \theta|\bar{\epsilon}_n + \sqrt{n}|\hat{\theta}^* - \theta|\bar{\epsilon}_n + r_n \underset{P_\theta}{\to} 0. ◄$$

Le théorème 1 montre que la méthode des approximations successives nous conduit en un pas, à partir de toute estimation asymptotiquement normale, dans un $o(1/\sqrt{n})$-voisinage de $\hat{\theta}^*$.

Si l'on exige l'existence des dérivées troisièmes continues $l'''(x, \theta)$, on peut partir de points plus éloignés de θ, disons d'une quantité $o(n^{-1/4})$. Dans ce cas, comme dans le théorème 1, en un pas on se trouve dans un $o(1/\sqrt{n})$-voisinage de $\hat{\theta}^*$). En effet

$$L'(X, t) = (t - \hat{\theta}^*)L''(X, \hat{\theta}^*) + \frac{(t - \hat{\theta}^*)^2}{2} L'''(X, \theta') =$$
$$= (t - \hat{\theta}^*)L''(X, t) + \frac{3}{2}(t - \hat{\theta}^*)^2 L''' - (X, \theta''),$$

où θ' et θ'' sont compris entre t et $\hat{\theta}^*$. Donc

$$U(\theta_n) - \hat{\theta}^* = \theta_n - \hat{\theta}^* - L'(X, \theta_n)(L''(X, \theta_n))^{-1} =$$
$$= \frac{3}{2}(\theta_n - \hat{\theta}^*)^2(I(\theta) + \varepsilon_n), \quad \sqrt{n}(U(\theta_n) - \hat{\theta}^*) \underset{P_\theta}{\to} 0,$$

si $|\theta_n - \theta| = o(n^{-1/4})$. ◄

EXEMPLE 1. *Classification des particules.* Considérons un émetteur de particules de deux types: des particules A avec la probabilité p et des particules B avec la probabilité $1 - p$. L'énergie des particules est aléatoire et admet une densité $f_1(x)$ pour les particules A et une densité $f_2(x)$ pour les particules B. Les fonctions $f_i(x)$ sont connues. On a enregistré n particules d'énergies respectives $x_1, \ldots, x_n$. On demande la probabilité p.

La fonction de vraisemblance vaut ici

$$f_p(X) = \prod_{i=1}^{n} (pf_1(x_i) + (1 - p)f_2(x_i)),$$

donc,

$$L'(X, p) = \sum_{i=1}^{n} \frac{f_1(x_i) - f_2(x_i)}{pf_1(x_i) + (1 - p)f_2(x_i)}. \tag{1}$$

Nous voyons que la recherche d'un estimateur du maximum de vraisemblance $\hat{p}^*$ nous conduit à une équation $L' = 0$ de degré $n - 1$ en p dont la résolution est très compliquée pour les grands n. Utilisons le théorème 1. A cet effet nous aurons besoin d'un estimateur asymptotiquement normal quelconque p^*. Supposons que $\int (F_1 - F_2)^2 dx \leqslant \infty$, où $F_i(x) = \int_{-\infty}^{x} f_i(t)dt$, et considérons l'approche suivante. Définissons p^* comme la valeur minimisant

$$\int (F_n^*(x) - F(x))^2 dx, \quad F(x) = pF_1(x) + (1 - p)F_2(x). \tag{2}$$

En égalant à 0 la dérivée de (2), on obtient $\int (F_n^* - F)(F_1 - F_2)dx = 0$,

$$p^* = \frac{\int (F_n^* - F_2)(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}.$$

Il est aisé de voir que $\mathbf{E}p^* = p$ et que

$$(p^* - p)\sqrt{n} = \frac{\int (F_n^* - F)\sqrt{n}(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}. \tag{3}$$

Des résultats des §§ 1.6 à 1.8 il s'ensuit que p^* est un estimateur asymptotiquement normal et que la distribution limite (3) est confondue avec la distribution

$$\frac{\int w^\circ(F(x))(F_1 - F_2)dx}{\int (F_1 - F_2)^2 dx}.$$

Donc, d'après le théorème 1, l'estimateur

$$p_1^* = p^* - L'(X, p^*)(L''(X, p^*))^{-1},$$

où L' est définie dans (1) et

$$L'' = -\sum \frac{(f_1(x_i) - f_2(x_i))^2}{(pf_1(x_i) + (1 - p)f_2(x_i))^2},$$

sera asymptotiquement équivalent à l'estimateur du maximum de vraisemblance $\hat{p}^*$. Le paramètre de dispersion de p_1^* sera défini par la quantité d'information

$$I(p) = \int \frac{(f_1(x) - f_2(x))^2}{pf_1(x) + (1 - p)f_2(x)} dx$$

et sera strictement inférieur à celui de p^*.

EXEMPLE 2. Nous proposons au lecteur de trouver de façon analogue une approximation pour l'estimateur du maximum de vraisemblance du paramètre α de la distribution de Cauchy $\mathbf{K}_{\alpha,1}$ de densité

$$k_{\alpha,1}(x) = \frac{1}{\pi(1 + (x - \alpha)^2)}.$$

Pour estimateur asymptotiquement normal « préliminaire » on peut prendre la médiane empirique ζ^* (cf. § 2 ou §§ 1.3, 1.8. L'estimateur $\alpha^* = \bar{x}$ est exclu, car $\mathbf{E}_\alpha \alpha^*$ n'existe pas). L'estimateur

$$\alpha_1^* = \zeta^* - L'(X, \zeta^*)(L''(X, \zeta^*))^{-1},$$

où

$$L'(X, \alpha) = -2 \sum \frac{x_i - \alpha}{1 + (x_i - \alpha)^2},$$

$$L''(X, \alpha) = 2 \sum \frac{1 - (x_i - \alpha)^2}{(1 + (x_i - \alpha)^2)^2},$$

sera asymptotiquement équivalent à l'estimateur du maximum de vraisemblance $\hat{\alpha}^*$. Comme

$$I(\alpha) = \int \frac{(k'_{\alpha,1}(x))^2}{k_{\alpha,1}(x)} dx = \frac{4}{\pi} \int \frac{x^2}{(1 + x^2)^3} dx = \frac{1}{2},$$

les paramètres de dispersion de ζ^* et de $\hat{\alpha}_1^*$ seront respectivement égaux à (cf. § 2)

$$\frac{1}{2k_{\alpha,1}(\alpha)} = \frac{\pi}{2}, \quad I^{-1/2}(\alpha) = \sqrt{2}, \quad \frac{\pi}{2} > \sqrt{2}.$$

EXEMPLE 3. Chaque être humain est de l'un des quatre groupes sanguins suivants : 0 (zéro), A, B, AB. La transmission du groupe sanguin est commandée par trois gènes : A, B et 0, le gène 0 étant dominé par les gènes A et B. Si donc p, q et $r = 1 - p - q$ représentent les probabilités d'apparition des gènes A, B et 0, les probabilités d'apparition des groupes sanguins seront égales aux quantités suivantes :

Tableau 1

i (numéro du groupe)	Groupe	Combinaisons donnant ce groupe	Probabilités
1	0	00	r^2
2	A	AA, A0	$p^2 + 2pr$
3	B	BB, B0	$q^2 + 2qr$
4	AB	AB	$2pq$

Soient ν_1, ν_2, ν_3, ν_4 les fréquences d'apparition des groupes respectifs dans une population de n individus. Comment calculer l'estimation du maximum de vraisemblance pour p et q? Les probabilités $p_i(\theta)$, $\theta = (p, q)$, d'apparition du i-ième groupe sanguin et leurs dérivées partielles par rapport à p et à q sont représentées dans le tableau 2.

Tableau 2

	i			
	1	2	3	4
$p_i(\theta)$	r^2	$p(p + 2r)$	$q(q + 2r)$	$2pq$
$\dfrac{\partial p_i(\theta)}{\partial p}$	$-2r$	$2r$	$-2q$	$2q$
$\dfrac{\partial p_i(\theta)}{\partial q}$	$-2r$	$-2p$	$2r$	$2p$

Pour le logarithme de la fonction de vraisemblance $L(X, \theta) = \sum_{i=1}^{4} \nu_i \ln p_i(\theta)$ on obtient donc

$$\frac{\partial L}{\partial p} = \sum \frac{\nu_i}{p_i} \frac{\partial p_i}{\partial p} = -\frac{2\nu_1}{r} + \frac{2r\nu_2}{p(p + 2r)} - \frac{2\nu_3}{q + 2r} + \frac{\nu_4}{p},$$

$$\frac{\partial L}{\partial q} = \sum \frac{\nu_i}{p_i} \frac{\partial p_i}{\partial p} = -\frac{2\nu_1}{r} - \frac{2\nu_2}{p + 2r} + \frac{2r\nu_3}{q(q + 2r)} + \frac{\nu_4}{q}. \qquad (4)$$

En égalant ces dérivées à zéro, on obtient pour θ^* un système de deux équations du quatrième degré. La résolution de ce système soulève de grosses difficultés techniques. Aussi est-il plus simple d'appliquer le théorème 1. A cet effet, on remarquera que

$$p_1 = r^2, \quad p_1 + p_2 = (p + r)^2, \quad p_1 + p_3 = (q + r)^2. \qquad (5)$$

Les estimations efficaces de p_i sont égales à $p_i^* = \nu_i/n$. En les portant dans (5) et en résolvant les équations obtenues, on trouve

$$p^* = \sqrt{p_1^* + p_2^*} - \sqrt{p_1^*}, \quad q^* = \sqrt{p_1^* + p_3^*} - \sqrt{p_1^*}.$$

Comme p_i^* est une estimation asymptotiquement normale de p_i (autrement dit, $(p_i^* - p_i)\sqrt{n} \in \Phi_{0, p_i(1 - p_i)}$), il en sera de même de p^* et q^* pour p et q en vertu des théorèmes du § 1.5.

Pour appliquer le théorème 1, il reste à calculer la matrice $(L''(X, \theta^*))^{-1}$ ou la matrice $(nI(\theta^*))^{-1}$, $\theta^* = (p^*, q^*)$.

Citons un exemple d'échantillon X obtenu par sondage d'une population de $n = 353$ individus.

Tableau 3

	0	A	B	AB	Total
ν_i p_i^*	121 0,343	120 0,340	79 0,224	33 0,093	353 1

Ce tableau nous donne $p^* = 0{,}241$, $q^* = 0{,}167$, $r^* = 1 - p^* - q^* = 0{,}592$. Le tableau 2 nous donne pour les éléments de la matrice $I(\theta)$, $\theta = \theta^*$,

$$\sum \left(\frac{\partial p_i(\theta)}{\partial p}\right)^2 \frac{1}{p_i(\theta)} = 4 + \frac{4r^2}{p(p + 2r)} + \frac{4q}{q + 2r} + \frac{2q}{p} = 9{,}970,$$

$$\sum \left(\frac{\partial p_i(\theta)}{\partial p}\right)^2 \frac{1}{p_i(\theta)} = 4 + \frac{4p}{p + 2r} + \frac{4r^2}{q(q + 2r)} + \frac{2p}{q} = 13{,}761,$$

$$\sum \frac{\partial p_i(\theta)}{\partial p} \cdot \frac{\partial p_i(\theta)}{\partial q} \cdot \frac{1}{p_i(\theta)} = 4 - \frac{4r}{p + 2r} - \frac{4r}{q + 2r} + 2 = 2{,}585.$$

D'où $|I(\theta^*)| = 130{,}512$,

$$I^{-1}(\theta^*) = \begin{vmatrix} 0{,}105 & -0{,}020 \\ -0{,}020 & 0{,}076 \end{vmatrix}.$$

En se servant des formules de $\dfrac{\partial L}{\partial p}$ et $\dfrac{\partial L}{\partial q}$ (cf. (4)), on trouve

$$L'(\theta^*, X) = (25{,}443,\ 34{,}161), \tag{6}$$

de sorte que pour la deuxième approximation de θ_1^* on a

$$\theta_1^* = \theta^* + \frac{1}{n} L'(\theta^*, X) I^{-1}(\theta^*) = (0{,}246,\ 0{,}173). \tag{7}$$

Ceci combiné au tableau 3 nous donne les estimations consignées dans le tableau 3A.

Tableau 3A

	0	A	B	AB
$p_i(\theta^*)$ $p_i(\theta_1^*)$	0,351 0,337	0,343 0,347	0,226 0,231	0,080 0,085

L'estimation θ_2^* ne sera pas modifiée par les itérations suivantes de la forme (7) (dans le cadre de la précision requise), puisque

$$L'(\theta_2^*, X) = (-0{,}076, \ -0{,}167)$$

(comparer avec (6)), de sorte que la troisième approximation de $\hat{\theta}^*$ et les suivantes seront confondues avec θ_2^*.

§ 27*. Propriétés des estimateurs du maximum de vraisemblance en l'absence des conditions de régularité. Convergence

Ce paragraphe, de même que le § 22, se tient à l'écart de l'exposé principal et traite le cas irrégulier. On se bornera à la démonstration de la convergence forte de l'estimateur du maximum de vraisemblance sous des conditions très faibles sur $f_t(x)$ et sans les conditions (RR) ou (R). Les propriétés de l'estimateur du maximum de vraisemblance et du rapport de vraisemblance dans le cas irrégulier font l'objet d'un examen plus détaillé dans [42].

Dans ce paragraphe on admettra que sont réalisées les conditions (A_μ), (A_c) et (A_0) et l'on désignera la distance de Kullback-Leibler $\varrho_1(\mathbf{P}_\theta, \mathbf{P}_t)$ par

$$\varrho(\theta, t) = \int \ln \frac{f_\theta(x)}{f_t(x)} \cdot f_\theta(x)\mu(dx).$$

On sait que si la condition (A_0) est remplie, $\varrho(\theta, t) > 0$ pour $t \neq \theta$.

Il est évident que la condition (A_0) est nécessaire à la convergence de l'estimateur du maximum de vraisemblance, c'est-à-dire à la convergence $\hat{\theta}^* \underset{P_\theta}{\to} \theta$. Si par exemple $\varrho(\theta, t_0) = 0$ pour $t_0 \neq \theta$, les points θ et t_0 seront tout simplement indiscernables ; les distributions $\mathbf{P}_\theta$ et $\mathbf{P}_{t_0}$ seront confondues et, si $X \in \mathbf{P}_\theta$ ou $X \in \mathbf{P}_{t_0}$, l'estimateur $\hat{\theta}^*$ ne peut être convergent quel que soit l'estimateur vers lequel il tend.

Il existe une variante uniforme $(\bar{A}_0)$ de la condition (A_0) (θ étant fixe) :

$(\bar{A}_0)$ *Pour tout* $\delta = \varepsilon(\delta) > 0$

$$\inf_{t\,:\,|t-\theta|\geq\delta} \varrho(\theta, t) > \varepsilon$$

pour un $\varepsilon > 0$.

Il est évident que $(\bar{A}_0)$ est une conséquence de (A_0), (A_c) et de la continuité de $\varrho(\theta, t)$. Donc, la condition $(\bar{A}_0)$ est aussi nécessaire.

Resserrons maintenant la condition $(\bar{A}_0)$. Posons

$$f_t^\Delta(x) = \sup_{|u|\leq\Delta} f_{t+u}(x).$$

(A_0^Δ) *Pour tout $\delta > 0$ il existe un $\Delta = \Delta(\delta) > 0$ tel que pour tous les t, $|t - \theta| > \delta$, et un $\epsilon > 0$, on a*

$$\int \ln \frac{f_t^\Delta(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < -\epsilon. \tag{1}$$

Cette condition est suffisante pour la convergence forte de l'estimateur du maximum de vraisemblance. Elle est proche de la condition $(\overline{A}_0)$ et de ce point de vue est proche d'une condition nécessaire. Mais la condition $(\overline{A}_0)$ est insuffisante à elle seule à la convergence de l'estimateur du maximum de vraisemblance (cf. remarque 1).

THÉORÈME 1. *Si la condition (A_0^Δ) est remplie, l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est fortement convergent.*

DÉMONSTRATION. L'estimation du maximum de vraisemblance $\hat{\theta}^*$ est un point t de maximum de la fonction $\psi(t, \theta, \mathbf{P}_n^*)$, où

$$\psi(\theta, t, \mathbf{P}) = \int \ln \frac{f_t(x)}{f_\theta(x)} \mathbf{P}(dx).$$

Comme $\psi(\theta, \hat{\theta}^*, \mathbf{P}_n^*) \geqslant \psi(\theta, \theta, \mathbf{P}_n^*) = 0$, pour prouver le théorème il suffit de s'assurer que

$$\operatorname{lim.sup}_{n \to \infty} \sup_{|t-\theta| \geqslant \delta} \psi(\theta, t, \mathbf{P}_n^*) < -\epsilon$$

$\mathbf{P}_\theta$-presque sûrement pour un certain $\varepsilon > 0$. (Ceci exprimera que $|\hat{\theta}^* - \theta| < \delta$ pour presque tous les $X_\infty \in \mathbf{P}_\theta$ à partir d'un certain $n = n(X_\infty) < \infty$.) Fixons δ et supposons que Δ vérifie la condition (1). Recouvrons l'ensemble $\Theta \setminus [\theta - \delta, \theta + \delta]$ avec les intervalles $\Delta_k = \{t : |t - t_k| \leqslant \Delta\}$, $k = 1, \ldots, N < \infty$, où $t_k \in \Theta$, $t_k \notin [\theta - \delta, \theta + \delta]$. La loi forte des grands nombres nous dit alors que

$$\sup_{|t-\theta| \geqslant \delta} \psi(\theta, t, \mathbf{P}_n^*) \leqslant \max_k \sup_{t \in \Delta_k} \psi(\theta, t, \mathbf{P}_n^*) \leqslant$$

$$\leqslant \max_k \frac{1}{n} \sum_{i=1}^n \sup_{t \in \Delta_k} \ln \frac{f_t(x_i)}{f_\theta(x_i)} \xrightarrow[\text{p.s}]{} \max_k \mathbf{E}_\theta \ln \frac{f_{t_k}^\Delta(x_1)}{f_\theta(x_1)} < -\epsilon. \blacktriangleleft$$

REMARQUE 1. Comme déjà signalé, la seule condition $(\overline{A}_0)$ est insuffisante à la convergence de $\hat{\theta}^*$. Pour nous en assurer considérons l'exemple suivant. Soient $\Theta = [0, 1]$, $\mathbf{P}_\theta = \mathbf{U}_{\theta, 1+\theta}$ pour $0 \leqslant \theta \leqslant 1/2$ et pour $\theta = 1$. Si $\theta \in]1/2, 1[$, la distribution $\mathbf{P}_\theta$ admet la densité $f_\theta(x) = 1/\theta$ pour $x \in]1 - \theta, 1[$. Supposons maintenant que $X \in \mathbf{P}_\theta = \mathbf{U}_{0,1}$. La condition $(\overline{A}_0)$ est alors remplie, puisque $\varrho(0, t) = -\infty$ pour $t \neq 0$. Dans le même temps il est immédiat de voir que $f_t(X) > 1$ pour $t \in]1 - x_{(1)}, 1[$ et que $\hat{\theta}^* = 1 - x_{(1)} \xrightarrow[\text{p.s}]{} 1$.

La condition (A_0^{Δ}) peut être représentée sous une forme équivalente légèrement différente. Désignons $f_t^{\circ}(x) = \lim\sup_{u \to t} f_u(x)$.

THÉORÈME 2. *La condition* (A_0^{Δ}) *est équivalente à la réalisation simultanée des deux conditions suivantes :*

(A_0°). *Pour tous les* $t \neq \theta$

$$\int \ln \frac{f_t^{\circ}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < 0.$$

(J). *Pour tous les* t *et un* $\Delta > 0$

$$\int \ln \frac{f_t^{\Delta}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < \infty.$$

La condition (J), de même d'ailleurs que les conditions (A_0°) et (A_0^{Δ}), exprime que les parties strictement positives des intégrants sont intégrables. De telles fonctions seront dites *intégrables supérieurement.*

En vertu de (A_c) la condition (J) est en fait équivalente à la majorabilité de l'intégrale

$$\int \ln \frac{f^{\Theta}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) < \infty, \tag{2}$$

où $f^{\Theta}(x) = \sup_{t \in \Theta} f_t(x)$,

DÉMONSTRATION du théorème 2. Il est évident que la condition (A_0^{Δ}) entraîne (A_0°) et (J). Supposons maintenant que sont remplies les conditions (A_0°) et (J). Si l'on admet que (A_0^{Δ}) est mise en défaut, on peut exhiber des suites $t_k \to t \in \Theta$, $\Delta_k \to 0$, $\varepsilon_k \to 0$ telles que

$$\int \ln \frac{f_{t_k}^{\Delta_k}(x)}{f_\theta(x)} \cdot f_\theta(x)\mu(dx) > -\varepsilon_k.$$

L'intégrant est majoré en vertu de la condition (J) par une fonction intégrable supérieurement, donc d'après le lemme de Fatou

$$\lim\sup_{k \to \infty} \int \ln \frac{f_{t_k}^{\Delta_k}(x)}{f_\theta(x)} f_\theta(x)\mu(dx) \leqslant \int \ln \frac{f_t^{\circ}(x)}{f_\theta(x)} f_\theta(x)\mu(dx) < 0.$$

Cette contradiction prouve le théorème. ◄

Indiquons maintenant des conditions suffisantes plus simples assurant la réalisation de (A_0°) et de (J), donc la convergence forte de l'estimateur du maximum de vraisemblance.

DÉFINITION 1. On dira que $f_t(x)$ *est de classe* D_0 si pour tout $t \in \Theta$ il existe un ensemble $C_t \in \mathfrak{B}_{\mathscr{X}}$, $\mathbf{P}_\theta(C_t) = 1$, sur lequel $f_t(x)$ est continue par rapport à t : $f_{t_k}(x) \to f_t(x)$ pour $t_k \to t$, $x \in C_t$.

Outre les fonctions $f_t(x)$ continues par rapport à t sur un ensemble C, $\mathbf{P}_\theta(C) = 1$, indépendant de t, sont de classe D_0 les fonctions dont les dérivées $f_t(x)$ possèdent dans le plan (t, x) des lignes de discontinuité isolées ne présentant pas de tronçons parallèles à l'axe des x. Telles sont notamment les fonctions dont les dérivées $f_t(x)$ présentent en tant que fonctions de x des discontinuités isolées aux points $x_t^{(1)}, x_t^{(2)}, \ldots$, dépendant continûment de t.

THÉORÈME 3. *Si $f_t(x) \in D_0$ et si est réalisée la condition (J), alors il en est de même de la condition $(A\mathfrak{F})$ et par suite l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est fortement convergent.*

DÉMONSTRATION. Si $f_t(x) \in D_0$, alors $f_t^\circ(x) = f_t(x)$ pour $x \in C_t$ et par suite

$$\int \ln \frac{f_t^\circ(x)}{f_\theta(x)} f_\theta(x)\mu(dx) = -\varrho(\theta, t) < 0. \quad ◄$$

COROLLAIRE 1. *Si $f_t(x) \in D_0$ est bornée et l'intégrale*

$$\int f_\theta(x) \ln f_\theta(x)\mu(dx) \tag{3}$$

finie, l'estimateur du maximum de vraisemblance est fortement convergent.

Le corollaire 1 résulte directement du théorème 3, puisque la majorabilité de $f_t(x)$ et la finitude de l'intégrale (3) entraînent la condition (J).

COROLLAIRE 2. *Si*

$$\varphi(\Delta) = \int \sup_{|u| < \Delta} |f_{t+u}(x) - f_t(x)|\mu(dx) \to 0 \tag{4}$$

lorsque $\Delta \to 0$, l'estimateur du maximum de vraisemblance est fortement convergent.

DÉMONSTRATION. Appliquons le théorème 3. Il est évident que $f_t(x)$ est de classe D_0, puisque la relation (4) ne peut être réalisée que dans le cas où $f_{t+u}(x) \to f_t(x)$ lorsque $u \to 0$ pour $[\mu]$-presque toutes les valeurs de x. D'autre part,

$$\int f_t^\Delta(x)\mu(dx) \leqslant \varphi(\Delta) + \int f_t(x)\mu(dx) = \varphi(\Delta) + 1,$$

et la condition (4) traduit également l'intégrabilité de $f_t^\Delta(x)$. Vu que $\ln \dfrac{f_t^\Delta(x)}{f_\theta(x)} \leqslant \dfrac{f_t^\Delta(x)}{f_\theta(x)} - 1$, on en déduit que l'intégrale de la condition (J) est

inférieure à

$$\int f_t^{\Delta}(x)\mu(dx) - 1 \leqslant \varphi(\Delta). \quad ◄$$

Au lieu de (4) on aurait pu exiger la convergence vers 0 de la quantité

$$\varphi_1(\Delta) = \int \sup_{|u| \leqslant \Delta} (\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)}\,)\,\mu(dx),$$

puisque $\varphi(\Delta)$ peut être majorée à l'aide de $\varphi_1(\Delta)$ comme suit

$$\varphi(\Delta) \leqslant \int \sup_{|u| \leqslant \Delta} |\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)}\,| \sup_{|u| \leqslant \Delta} |\sqrt{f_{t+u}(x)} + \sqrt{f_t(x)}\,|\mu(dx) \leqslant$$

$$\leqslant \varphi_1^{1/2}(\Delta)\left\{\int \sup_{|u| \leqslant \Delta} (\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)} + 2\sqrt{f_t(x)}\,)^2\mu(dx)\right\}^{1/2} \leqslant$$

$$\leqslant [2\varphi_1(\Delta)(\varphi_1(\Delta + 4)]^{1/2}.$$

COROLLAIRE 3. *Si $f_t(x)$ est dérivable par rapport à t pour $[\mu]$-presque toutes les valeurs de x et si*

$$\int |f_t'(x)|\mu(dx) < c < \infty, \tag{5}$$

l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est fortement convergent. La condition (5) *est toujours remplie si la quantité d'information de Fisher $I(t)$ est bornée.*

Nous sommes arrivés au résultat que nous aurions pu obtenir en appliquant le théorème 23.2. Le procédé de démonstration de ce dernier (cf. §§ 21, 23) montre que la majorabilité de $I(t)$ ou de (5) n'est pas essentielle pour le corollaire 3 si la distance de Hellinger $\varrho_3(\mathbf{P}_\theta, \mathbf{P}_{\theta+\Delta})$ est uniformément différente de zéro pour $|\Delta| \geqslant \delta > 0$.

DÉMONSTRATION. Il est évident que $f_t(x)$ est de classe D_0. Pour que la condition (J) soit réalisée, il suffit, comme nous l'avons vu dans la démonstration du corollaire 2, que $f_t^{\Delta}(x)$ soit intégrable. Mais

$$\int f_t^{\Delta}(x)\mu(dx) \leqslant \int \left[f_t(x) + \int_{-\Delta}^{\Delta} |f_{t+u}'(x)|du\right]\mu(dx) =$$

$$= 1 + \int_{-\Delta}^{\Delta}\left[\int |f_{t+u}'(x)|\mu(dx)\right]du \leqslant 1 + 2\Delta c.$$

Reste à appliquer le théorème 3. La dernière proposition du corollaire 3 résulte de l'inégalité de Cauchy-Bouniakovski, puisqu'en vertu de cette dernière

$$\int |f_t'(x)|\mu(dx) \leqslant I^{1/2}(t). \quad ◄$$

COROLLAIRE 4. *Supposons que θ est le paramètre de translation de la famille $f_\theta(x) = f(x - \theta)$, $\int f(x)\ln f(x)dx > -\infty$. Si la fonction $f(x)$ est bornée* (*sinon la méthode du maximum de vraisemblance perd son sens* (cf. § 26)) *et présente un ensemble B de points de discontinuité, dont la mesure de Lebesgue $\mu(\overline{B}^c)$ de l'adhérence est nulle, alors l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est fortement convergent.*

DÉMONSTRATION. Assurons-nous que les conditions du théorème 3 sont remplies. La condition (J) est réalisée de façon évidente. La relation $f_t(x) \in D_0$ résulte de la définition de D_0 dans laquelle il faut poser $C_t = \overline{B}^c - t$ (ceci est une translation de vecteur t de l'ensemble $\overline{B}^c$, complémentaire de l'adhérence de B). L'ensemble $\overline{B}^c$ étant ouvert, la relation $x - t \in \overline{B}^c - t$ entraîne $x - t_k \in \overline{B}^c - t$ pour les $|t_k - t|$ assez petits. Ceci exprime que $f(x - t_k) \to f(x - t)$. ◄

Signalons qu'il est superflue de supposer que la condition (A_0) est réalisée dans le corollaire 4, car elle l'est automatiquement. Si l'on admet que ($A8$) n'est pas remplie, on arrive à une fonction $f(x)$ périodique, ce qui est impossible.

S'agissant des conditions du corollaire 4 on remarquera que la condition portant sur la « continuité » de $f(x)$ est assez faible. Elle non plus n'est visiblement pas essentielle. C'est ce qu'indique dans une certaine mesure l'exemple suivant.

EXEMPLE 1. Soit $f(x)$ une fonction arbitraire à support borné $]a, b[= \{x : f(x) > 0\}$. Alors

$$\mathbf{P}_\theta(|\hat{\theta}^* - \theta| > \delta) \leqslant (1 - F_0(a + \delta))^n + F_0^n(b - \delta), \tag{6}$$

où $F_\theta(x) = \int_{-\infty}^{x} f_\theta(y)dy$. L'inégalité (6) exprime la convergence forte de $\hat{\theta}^*$.

Elle résulte des relations

$$\{\hat{\theta}^* - \theta > \delta\} \subset \left\{\prod_{i=1}^{n} f_{\theta+\delta}(x_i) > 0\right\} \subset \bigcap_{i=1}^{n} \{x_i \geqslant a + \theta + \delta\},$$

$$\mathbf{P}_\theta(\hat{\theta}^* - \theta > \delta) \leqslant [1 - F_\theta(a + \theta + \delta)]^n = [1 - F_0(a + \delta)]^n.$$

La condition portant sur la finitude de l'intégrale $\int f(x)\ln f(x)dx$ dans le corollaire 4 n'est pas non plus essentielle : on peut exhiber un exemple dans lequel cette intégrale prend la valeur $-\infty$, alors que la condition (J) est remplie.

Des remarques du § 18 il s'ensuit que tout ce qui a été dit dans et après le corollaire 4 reste entièrement en vigueur pour le paramètre d'échelle.

§ 28. Les résultats des §§ 23 à 27 pour un paramètre vectoriel

Dans ce paragraphe on généralise les principaux résultats des §§ 23 à 27 au cas vectoriel. On exposera ces résultats dans la même chronologie et l'on ne s'attardera que sur les points où l'introduction d'un paramètre vectoriel modifie soit l'énoncé du résultat soit les raisonnements.

Soit donc $\theta \in \Theta \subset R^k$, $k > 1$. Les énoncés des conditions (A_μ), (A_c), (A_0), de même que les définitions du rapport de vraisemblance

$$Z(u) = \frac{f_{\theta+u}(X)}{f_\theta(X)}$$

et de la distance de Hellinger

$$r(u) = \varrho(\mathbf{P}_{\theta+u}, \mathbf{P}_\theta) = \int(\sqrt{f_{\theta+u}(x)} - \sqrt{f_\theta(x)})^2\mu(dx),$$

ne sont pas liés à la dimension.

1. Inégalités pour le rapport de vraisemblance (résultats du § 23). Pour étudier le comportement de la fonction $Z(u)$ au voisinage de 0 nous aurons besoin de la condition suivante : *la fonction* $\sqrt{f_\theta(x)}$ *est dérivable par rapport à* θ, *la matrice d'information de Fisher*

$$I(\theta) = |I_{ij}(\theta)| = \left\| \mathbf{E}_\theta \frac{\partial}{\partial\theta_i} l(x_1, \theta) \frac{\partial}{\partial\theta_j} l(x_1, \theta) \right\| \qquad (1)$$

est bornée et définie positive pour tous les $\theta \in \Theta$.

Dans cette condition, le théorème 21.3A nous dit que pour tous les θ

$$0 < g \leqslant \frac{r(u)}{|u|^2} \leqslant h = \frac{1}{4}\sup_\theta \operatorname{Tr} I(\theta) < \infty. \qquad (2)$$

Ici et dans la suite $|u| = \sqrt{u_1^2 + \ldots + u_k^2}$ représente la norme euclidienne du vecteur $u = (u_1, \ldots, u_k)$.

La première proposition du théorème 23.1 et sa démonstration se généralisent au cas multidimensionnel sans changement, car elles ne sont pas liées à la dimension.

THÉORÈME 1. *Si* (2) *est remplie, on a*

$$\mathbf{E}_\theta Z^{1/2}(u) \leqslant e^{-ng|u|^2/2}.$$

Pour généraliser le théorème 23.2 nous aurons besoin de la condition subsidiaire suivante :

$$\gamma = \sup_\theta \mathbf{E}_\theta |l'(x_1, \theta)|^s < \infty \qquad (3)$$

pour un $s > k$.

THÉORÈME 2 (analogue du théorème 23.2). *Si les conditions* (2) *et* (3) *sont réalisées, alors pour tous* z, $n \geqslant 1$,

$$\mathbf{P}_\theta(\sup_{|v|>u} Z(v/\sqrt{n}) > e^z) \leqslant c\gamma e^{-z/2} e^{-\beta u^2}, \tag{4}$$

où $c < \infty$ *et* $\beta > 0$ *ne dépendent que de* k, g *et* s.

La démonstration, particulièrement simple du théorème 23.2, ne se généralise malheureusement pas au cas multidimensionnel. Ceci est dû au fait que le maximum de la fonction $p(u)$ dans le domaine $D \subset R^k$, $k >$ > 1, ne peut, contrairement au cas scalaire, être estimé par une intégrale de $|p'(u)|$ le long d'une courbe fixe de D. La nouvelle démonstration occupe beaucoup de place et nous avons jugé plus raisonnable de la proposer dans l'Annexe VII.

La démonstration des propositions relatives à la convergence de l'estimateur du maximum de vraisemblance et à l'estimation des moments du n° 2, § 23 n'est pas liée à la dimension. Ces propositions restent en vigueur sous la forme suivante.

THÉORÈME 3 (analogue du théorème 23.3). *Si les conditions* (2) *et* (4) *sont réunies, la relation* (23.6) *dans laquelle il faut remplacer* $g/4$ *par* β (cf. théorème 2) *est valable pour tous* z, $n \geqslant 1$.

Les propositions des corollaires 23.1 et 23.2 restent valables si l'on remplace encore $g/4$ par β.

2. Propriétés asymptotiques du rapport de vraisemblance (résultats du § 24). Par conditions (RR) dans le cas vectoriel on comprendra l'ensemble de conditions suivantes :

1) *Les conditions* (A_0), (A_c), (R).

2) *La fonction* $l(x, t)$ *est deux fois continûment dérivable par rapport à* θ *à l'intérieur de* Θ *pour* $[\mu]$*-presque toutes les valeurs de* x. *Ceci étant, on suppose que les dérivées*

$$l''_{ij}(x, t) = \frac{\partial^2 l(x, t)}{\partial t_i \partial t_j}$$

admettent un majorant $l(x)$ *indépendant de* t ($|l''_{ij}(x, t)| \leqslant l(x)$) *pour lequel l'intégrale*

$$\mathbf{E}_t l(x_1) = \int l(x) f_t(x) \mu(dx)$$

converge uniformément *) *en* $t \in \Theta$.

---

*) Cf. note de la page 221 sur la convergence uniforme du § 24.

3) *De plus on admettra en cas de besoin que la condition* (3) *est remplie.*

Comme en dimension un, nous aurons besoin des deux propriétés suivantes qui découlent des conditions (*RR*) :

1) *La possibilité d'une double dérivation par rapport à θ sous de signe d'intégration dans l'égalité*

$$\int f_\theta(x)\mu(dx) = 1,$$

possibilité qui exprime que

$$\int \frac{\partial}{\partial\theta_i} f_\theta(x)\mu(dx) = 0, \int \frac{\partial^2}{\partial\theta_i\partial\theta_j} f_\theta(x)\mu(dx) = 0. \tag{5}$$

2) *La convergence uniforme de l'intégrale $I(\theta)$* :

$$\sup_\theta \mathbf{E}_\theta[(l'(x_1, \theta))^2; \mid l'(x_1, \theta)| > N] \to 0 \tag{6}$$

lorsque $N \to \infty$.

La démonstration de ces propriétés a été reportée à l'Annexe VI. Pour alléger l'exposé on peut les inclure dans les conditions (*RR*).

D'après les égalités

$$l_i'(x, \theta) = \frac{1}{f_\theta(x)} \cdot \frac{\partial f_\theta(x)}{\partial\theta_i},$$

$$l_{ij}''(x, \theta) = \frac{1}{f_\theta(x)} \cdot \frac{\partial^2 f_\theta(x)}{\partial\theta_i\partial\theta_j} - \frac{1}{f_\theta^2(x)} \cdot \frac{\partial f_\theta(x)}{\partial\theta_i} \cdot \frac{\partial f_\theta(x)}{\partial\theta_j},$$

on déduit des relations (5) que

$$\mathbf{E}_\theta l_i'(x_1, \theta) = 0,$$

$$\mathbf{E}_\theta l_{ij}''(x_1, \theta) = -\mathbf{E}_\theta l_i'(x_1, \theta) l_j'(x_1, \theta) = -I_{ij}(\theta).$$

Comme en dimension un, les conditions (*RR*) *expriment que les théorèmes du* § *23 seront valables pour* $\sup_{|v| \geq u} Z(v/\sqrt{n})$ *et pour* $\sqrt{n}(\hat\theta^* - \theta)$.

Si les conditions (*RR*) sont remplies, on a les analogues suivants des lemmes 24.1 et 24.2.

LEMME 1. *Les fonctions $l_{ij}''(x, \theta)$ sont continues « en moyenne »* :

$$\mathbf{E}_\theta\omega_\Delta''(x_1) \to 0$$

uniformément par rapport à θ lorsque $\Delta \to 0$, où $\omega_\Delta''(x) = \max_{i,j}.\ \sup_{\theta, |u| - \Delta} |l_{ij}''(x, \theta + u) - l_{ij}''(x, \theta)|$.

La démonstration reprend mot à mot celle du lemme 24.1. ◄

Posons

$$\gamma_n(\delta, \theta) = \sup_{\substack{\Delta \leqslant \delta \\ |\omega| = 1}} \left| \frac{(L'(X, \theta + \omega\Delta), \omega) - (L'(X, \theta), \omega)}{n\Delta} + \omega I(\theta)\omega^T \right|.$$

LEMME 2 (analogue du lemme 24.2). *Supposons remplies les conditions* (RR) *et soit* $\delta_n > 0$ *une suite convergeant vers* 0. *Alors pour* $X \in \mathbf{P}_\theta$

$$\gamma_n(\delta_n, \theta) \underset{\text{p.s.}}{\to} 0, \quad \gamma_n(\delta_n, \hat{\theta}^*) \underset{\text{p.s.}}{\to} 0.$$

Dans ces relations les valeurs $I(\theta)$ *et* $I(\hat{\theta}^*)$ *peuvent être substituées l'une à l'autre.*

DÉMONSTRATION. Comme en dimension un, il nous suffit de nous assurer que $\gamma_n(\delta_n) \to 0$, où

$$\gamma_n(\delta) = \sup_{\substack{\Delta \leqslant \delta \\ |\omega| = 1}} \left| \frac{(L'(X, \theta + \omega\Delta), \omega) - (L'(X, \theta), u)}{n\Delta} - \frac{\omega L''(X, \theta)\omega^T}{n} \right|.$$

Or, $\gamma_n(\delta_n) \leqslant \frac{1}{n} \sum_i \sum_{k,j} \omega_{\delta_n}''(x_i)|\omega_k\omega_j|$, où $\omega_\delta''(x)$ est le plus grand module de continuité des fonctions $l_{ij}''(x, \theta)$. Comme

$$\sum_{k,j} |\omega_k\omega_j| \leqslant k|\omega|^2 = k,$$

il vient

$$\gamma_n(\delta_n) \leqslant \frac{k}{n} \sum_i \omega_{\delta_n}''(x_i). \tag{7}$$

La suite de la démonstration repose sur le lemme 1 et reprend exactement les raisonnements du lemme 24.2. ◄

Le théorème 24.1 se généralise au cas multidimensionnel de la manière suivante.

THÉORÈME 4. *Supposons remplies les conditions* (RR) *et soit* $\delta_n > 0$, $n = 1, 2, \ldots$, *une suite convergeant vers* 0. *Si* $X \in \mathbf{P}_\theta$, *alors pour les* u *tels que* $|u/\sqrt{n}| \leqslant \delta_n$, *on a*

$$Y(u) \equiv \ln Z(u/\sqrt{n}) = (\xi_n, u) - \frac{1}{2} uI(\theta)u^T(1 + \epsilon_n(X, \theta, u)), \tag{8}$$

où

$$|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta) \underset{\text{p.s.}}{\to} 0,$$

$$\xi_n = \frac{1}{\sqrt{n}} \operatorname{grad} L(X, \theta) = \frac{1}{\sqrt{n}} L'(X, \theta) \in \Phi_{0, I(\theta)}.$$

La valeur $u^* = \sqrt{n}\,(\hat{\theta}^* - \theta)$ qui réalise le maximum de $Y(u)$ se représente sous la forme

$$u^* = \xi_n I^{-1}(\theta)(E + \epsilon_n(X, \theta)), \quad \epsilon_n(X, \theta) \underset{\text{p.s.}}{\to} 0, \tag{9}$$

où E est la matrice unité. En outre

$$2Y(u^*) = \xi_n I^{-1}(\theta)\xi_n^T(1 + \epsilon_n(X, \theta)) \in$$
$$\in \frac{1}{2}\,\xi I^{-1}(\theta)\xi^T \in \mathbf{H}_k, \ \xi \in \Phi_{0, I(\theta)}. \tag{10}$$

Parallèlement à (8) *on a la représentation*

$$Y(u) - Y(u^*) = \frac{1}{2}(u - u^*)I(\theta)(u - u^*)^T(1 + \epsilon_n(X, \theta, u)),$$
$$|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta).$$

Dans toutes les assertions énoncées on peut remplacer $I(\theta)$ par $I(\hat{\theta}^*)$.

De même que dans le § 24, par $\epsilon_n(X, \theta)$ on comprend ici des suites convergeant presque sûrement vers 0 par rapport à $\mathbf{P}_\theta$.

A noter encore que la partie principale de (8) peut être écrite sous la forme

$$\xi_n u^T - \frac{1}{2}\,uI(\theta)u^T =$$
$$= -\frac{1}{2}(u - \xi_n I^{-1}(\theta))I(\theta)(u - \xi_n I^{-1}(\theta))^T + \frac{1}{2}\,\xi_n I^{-1}(\theta)\xi_n^T.$$

Ceci représente la densité d'une loi normale multidimensionnelle de moyenne $\xi_n I^{-1}(\theta)$ et de matrice des moments d'ordre deux $I^{-1}(\theta)$.

DÉMONSTRATION du théorème 4. Elle reprend *ad litteram* celle du théorème 24.1 Pour $\Delta \leqslant \delta_n$ le lemme 2 nous donne

$$(L'(X, \theta + \Delta\omega), \omega) = (L'(X, \theta), \omega) - n\Delta\omega I(\theta)\omega^T(1 + \epsilon_n(X, \theta, \Delta\omega)),$$
$$|\epsilon_n(X, \theta, \Delta\omega)| \leqslant \epsilon_n(X, \theta).$$

En intégrant cette égalité par rapport à Δ entre 0 et $|u|/\sqrt{n}$ et en posant $\omega = u/|u|$, on obtient

$$L(X, \theta + u/\sqrt{n}) - L(X, \theta) = \int_0^{|u|/\sqrt{n}} (L'(X, \theta + \Delta u), u)d\Delta =$$
$$= \frac{|u|}{\sqrt{n}}(L'(X, \theta\ \omega) - \frac{|u|^2}{2}\,\omega I(\theta)\omega^T(1 + \epsilon_n(X, \theta, u)) =$$
$$= (\xi_n, u) - \frac{1}{2}\,uI(\theta)u^T(1 + \epsilon_u(X, \theta, u)), \quad |\epsilon_n(X, \theta, u)| \leqslant \epsilon_n(X, \theta).$$

Le théorème limite central multidimensionnel (cf. Annexe V) nous dit que

$$\xi_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l'(x_i, \theta) \in \Phi_{0,I(\theta)}.$$

Ce qui prouve la représentation (8). Les autres assertions du théorème se prouvent exactement comme dans le théorème 24.1 aux changements près liés au passage au cas multidimensionnel. La relation

$$\frac{1}{2} \xi I^{-1}(\theta)\xi^T \in \mathbf{H}_k$$

de (10) résulte des propriétés de la distribution normale (cf. n°4 du § 2.2). ◄

A propos de la relation (10) il est utile de faire la remarque suivante.

Remarque 1. Les matrices $I^{-1}(\theta)$ et $I(\theta)$ sont définies positives et il existe une matrice $I^{-1/2}(\theta)$ qui est racine carrée de la matrice $I^{-1}(\theta)$, c'est-à-dire vérifie la relation

$$I^{-1/2}(\theta)I^{-1/2}(\theta) = I^{-1}(\theta).$$

En effet, si une matrice $M > 0$ (i.e. définie positive), il existe une matrice orthogonale C pour laquelle $CMC^T = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ est une matrice diagonale dont les éléments diagonaux λ_i sont > 0. Si l'on pose maintenant $M^{1/2} = C^T \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_k^{1/2})C$, on obtient de toute évidence la racine carrée de M.

Ceci et la symétrie de la matrice $I^{-1}(\theta)$ nous permettent de mettre (10) sous la forme

$$\frac{1}{2} (\xi_n I^{-1/2}(\theta))(\xi_n I^{-1/2}(\theta))^T.$$

Le vecteur $\eta_n = \xi_n I^{-1/2}(\theta)$ est de toute évidence la somme normée de n vecteurs aléatoires indépendants équidistribués de moyenne nulle et de matrice des moments d'ordre deux

$$\mathbf{E}_\theta(\xi_n I^{-1/2}(\theta))^T(\xi_n I^{-1/2}(\theta)) = \mathbf{E}_\theta I^{-1/2}(\theta)\xi_n^T\xi_n I^{-1/2}(\theta) = E,$$

puisque

$$\mathbf{E}_\theta \xi_n^T \xi_n = \mathbf{E}_\theta(l'(x_1, \theta))^T(l'(x_1, \theta)) = I(\theta).$$

Ceci exprime que $\xi_n I^{-1/2}(\theta) \in \Phi_{0,E}$ d'après le théorème limite central multidimensionnel.

Théorème 5 (analogue du théorème 24.2). *Supposons remplies les conditions du théorème* 24.2 *pour* $\theta \in R^k$ *et* $\alpha = \beta/2$ *(*β *a été défini dans le théorème* 2). *Alors*

$$J \equiv \int w(u^* - u)q(\theta + u/\sqrt{n})Z(u/\sqrt{n})\Pi(du) = e^{Y(u^*)}q(\theta) \times$$
$$\times \left[\int w(u^* - u)\exp\left\{-\frac{1}{2}(u - u^*)I(\theta)(u - u^*)^T\right\}\Pi(du) + \epsilon_n(X, \theta)\right]. \quad (11)$$

Si Π *est la mesure de Lebesgue,* $\Pi(du) = du$, *alors*

$$J = \frac{(2\pi)^{k/2}}{\sqrt{|I(\theta)|}} e^{Y(u^*)}q(\theta)(\mathbf{E}w(\eta) + \epsilon_n(X, \theta)), \quad (12)$$

où $\epsilon_n(X, \theta) \underset{p.s.}{\to} 0$, $\eta \in \Phi_{0,I^{-1}(\theta)}$ *($\varepsilon_n(X, \theta)$ est une suite de vecteurs si $w(t)$ est une fonction vectorielle).*

La démonstration du théorème 5 est calquée sur celle du théorème 24.2, puisque cette dernière n'est pas liée à la dimension.

3. Propriétés de l'estimateur du maximum de vraisemblance (résultats du § 25). Nous admettrons partout dans ce numéro que les conditions (RR) sont remplies.

L'analogue du théorème 25.1 est de la forme suivante.

THÉORÈME 6. *L'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est asymptotiquement normal et de plus la convergence*

$$u^* = (\hat{\theta}^* - \theta)\sqrt{n} \Leftarrow \Phi_{0,I^{-1}(\theta)}$$

est réalisée simultanément pour les moments de tout ordre. En particulier,

$$\mathbf{E}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) \to I^{-1}(\theta). \tag{13}$$

Par ailleurs, pour toute fonction continue $w(t)$ telle que $|w(t)| < e^{\beta|t|^2/2}$ (le nombre β est défini dans le théorème 2), on a

$$\mathbf{E}_\theta w(u^*) \to \mathbf{E}w(\eta), \quad \eta \in \Phi_{0,I^{-1}(\theta)}.$$

La relation (13) exprime que $\hat{\theta}^* \in K_{\Phi,2}$.

Le théorème 6 résulte du théorème 4 (cf. (9)) et de l'analogue multidimensionnel du corollaire 23.2 qui découle à son tour du théorème 3 (comparer avec la démonstration du théorème 25.1). ◄

Définissons la classe $\bar{K}_0$ comme l'ensemble des estimateurs θ^* dont le biais $b(\theta) = (b_1(\theta), \ldots, b_k(\theta)) = \mathbf{E}_\theta\theta^* - \theta$ est tel que

$$|b(\theta)| = o(1/\sqrt{n}), \quad b_{ij}(\theta) = \frac{\partial b_i(\theta)}{\partial \theta_j} \to 0$$

lorsque $n \to \infty$.

Les analogues des théorèmes 25.2 et 25.3 sont de la même forme ici.

THÉORÈME 7. *L'estimateur $\hat{\theta}^*$ est un estimateur asymptotiquement R-efficace. De plus, $\hat{\theta}^* \in \bar{K}_0$ et est asymptotiquement efficace dans $\bar{K}_0$.*

La R-efficacité asymptotique de $\hat{\theta}^*$, qui est équivalente à (13), a visiblement lieu. La démonstration de l'appartenance de $\hat{\theta}^*$ à $\bar{K}_0$ et de l'efficacité asymptotique de $\hat{\theta}^*$ dans $\bar{K}_0$ s'effectue exactement comme en dimension un.

Passons maintenant à la propriété de bayésienneté asymptotique. Dire qu'un estimateur θ^* est asymptotiquement R-bayésien revient à dire par définition que (comparer avec le § 20)

$$\mathbf{E}(\theta^* - \theta)^T(\theta^* - \theta) = J/n + o(1/n), \quad J = \int I^{-1}(t)\mathbf{Q}(dt). \tag{14}$$

Dire que θ^* est asymptotiquement bayésien revient à dire que

$$\limsup_{n\to\infty}[nv(\theta^*) - nv(\theta^*_{\mathbf{Q}})] \leqslant 0, \tag{15}$$

où $\theta^*_{\mathbf{Q}}$ est un estimateur bayésien minimisant $v(\theta^*) = \mathbf{E}(\theta^* - \theta)V(\theta^* - \theta)^T$ pour toute matrice semi-définie positive V.

THÉORÈME 8 (analogue du théorème 25.4). *L'estimateur $\hat{\theta}^*$ est asymptotiquement R-bayésien. Si la distribution a priori* **Q** *admet une densité par rapport à la mesure de Lebesgue sur Θ, alors $\hat{\theta}^*$ est un estimateur asymptotiquement bayésien.*

DÉMONSTRATION. Elle est identique à celle du théorème 25.4. La relation (14) pour $\theta^* = \hat{\theta}^*$ résulte de ce que

$$\lim_{n\to\infty} \mathbf{E}n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) =$$
$$= \mathbf{E}\lim_{n\to\infty}\mathbf{E}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta) = \mathbf{E}I^{-1}(\theta) = J.$$

Le passage à la limite sous le signe de l'espérance mathématique (c'est-à-dire d'intégration) est licite, puisque la quantité $\mathbf{E}_\theta n(\hat{\theta}^* - \theta)^T(\hat{\theta}^* - \theta)$ est majorée par une constante indépendante de n et de θ (comparer avec le corollaire 23.2).

Pour prouver (15) on remarquera qu'en vertu des résultats du § 20 l'inégalité intégrale de Rao-Cramer dans le cas où **Q** admet une densité est de la forme

$$\mathbf{E}n(\theta^* - \theta)^T(\theta^* - \theta) \geqslant J + o(1).$$

Ceci exprime que

$$nv(\theta^*_{\mathbf{Q}}) \geqslant \sum v_{ij}J_{ij} + o(1),$$

où $|J_{ij}| = J$, $|v_{ij}| = V$. D'autre part, en vertu de (14) on a pour $\theta^* = \hat{\theta}^*$

$$nv(\hat{\theta}^*) = \sum v_{ij}J_{ij} + o(1).$$

De ces relations on déduit (15) pour $\theta^* = \hat{\theta}^*$. ◄

Les théorèmes 25.5 et 25.6 admettent aussi des analogues. Du théorème 5, par exemple, on déduit le

THÉORÈME 9 (analogue du théorème 25.6). *Soit* $X \in \mathbf{P}_\theta$ *et soit* θ *un point intérieur de* Θ. *Si* $q(t)$ *est une densité continue et strictement positive à l'intérieur de* Θ *d'une distribution a priori, alors*

$$\sqrt{n}(\hat{\theta}^* - \theta_Q^*) \underset{\text{p.s.}}{\to} 0,$$

où θ_Q^* *est un estimateur bayésien associé à* $q(t)$.

La minimaximalité asymptotique de $\hat{\theta}^*$ peut être établie, comme le théorème 25.7, à l'aide de l'analogue multidimensionnel du critère de minimaximalité asymptotique démontré dans le corollaire 20.3 :

$$\lim_{n\to\infty} \sup_{t\in\Gamma} \mathbf{E}_t n(\hat{\theta}^* - \theta) V(\hat{\theta}^* - \theta)^T = \sup_{t\in\Gamma} \sum I_{ij}^{-1}(\theta) v_{ij},$$

$$|I_{ij}^{-1}|(\theta)| = I^{-1}(\theta),$$

et à l'aide de la convergence uniforme dans (13) qui découle des résultats du paragraphe suivant.

Les propriétés d'optimalité asymptotique de $\hat{\theta}^*$ doivent être utilisées avec circonspection dans le cas où la dimension k du paramètre θ est élevée. Il faut veiller à ce que le rapport n/k (le nombre d'observations sur un paramètre scalaire) soit assez grand, sinon les conclusions risqueraient d'être fausses.

EXEMPLE 1. Les concentrations $\mu_1, \ldots, \mu_n$ de n solutions sont testées deux fois en laboratoire. On admet que la variance σ^2 des n observations $(x_1, y_1), \ldots, (x_n, y_n)$ est la même et que ces observations sont indépendantes et normales. On a donc

$$f_\theta(X) = \frac{1}{\sigma^{2n}(2\pi)^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum [(x_i - \mu_i)^2 + (y_i - \mu_i)^2] \right\},$$

où

$$\theta = (\mu_1, \ldots, \mu_n, \sigma^2).$$

Les estimateurs du maximum de vraisemblance de μ_i sont égaux à

$$\hat{\mu}_i^* = \frac{1}{2}(x_i + y_i).$$

Il est évident que ces estimateurs sont sans biais et ne convergent pas. L'estimateur du maximum de vraisemblance de σ^2 vaut

$$(\hat{\sigma^2})^* = \frac{1}{4n} \sum (x_i - y_i)^2 \underset{P}{\to} \sigma^2/2 \text{ lorsque } n \to \infty.$$

Cet estimateur donne avec une grande crédibilité une valeur fausse du paramètre σ^2 (une valeur deux fois moindre).

4. Calcul approché de l'estimateur du maximum de vraisemblance. Le § 26 reste entièrement en vigueur dans le cas multidimensionnel si l'on comprend par $[L''(X, t)]^{-1}$ la matrice inverse de $L''(X, t)$.

5. Propriétés de l'estimateur du maximum de vraisemblance en l'absence des conditions de régularité (résultats du § 27). Les conditions de convergence de θ formulées dans les théorèmes 27.1 à 27.3 sont indépendantes de la dimension. La démonstration de ces théorèmes reste entièrement en vigueur aux changements évidents près liés au fait qu'il faut recouvrir l'ensemble Θ (en vertu de la condition (A_c)) non plus par un nombre fini d'intervalles mais de boules. On peut en dire autant des corollaires 27.1 à 27.4.

§ 29. Uniformité en θ des propriétés asymptotiques du rapport de vraisemblance et des estimateurs du maximum de vraisemblance

Les propositions des §§ 24, 25, 28 dans leur version uniforme nous seront utiles dans la suite et essentiellement dans les §§ 13, 14 et 15 du prochain chapitre. La plupart de ces propositions (notamment les propositions relatives à la $\mathbf{P}_\theta$-distribution limite de $(\hat{\theta}^* - \theta)\sqrt{n}$) ont été établies dans l'hypothèse où θ est un point fixe de Θ. Voyons ce qui se passe lorsque θ varie avec n. Il est clair que la distribution $\mathbf{P}_\theta$ variera aussi, de sorte que chaque échantillon X_n aura sa « propre » distribution pour $n = 1, 2, \ldots$

Nous arrivons ainsi à un schéma de séries (cf. [11]) pour lequel les théorèmes limites fondamentaux seront d'une forme légèrement différente. En particulier, la loi forte des grands nombres n'a plus de sens, puisque les variables aléatoires envisagées ne sont plus définies (pour des n différents) sur un même espace probabilisé.

1. Loi des grands nombres et théorème limite central uniformes. Soient $X \in \mathbf{P}_\theta$, $\eta_{n,\theta} = \eta_n(X, \theta)$.

DÉFINITION 1. On dira qu'une suite $\eta_{n,\theta}$ *converge uniformément en probabilité vers une constante* $a(\theta)$ si pour tout $\varepsilon > 0$

$$\sup_{\theta \in \Theta} \mathbf{P}_\theta(|\eta_{n,\theta} - a(\theta)| > \varepsilon) \to 0$$

lorsque $n \to \infty$.

Nous écrirons cette relation sous la forme « $\eta_{n,\theta} \xrightarrow[P_\theta]{} a(\theta)$ *uniformément en* θ ».

DÉFINITION 2. On dira que $\eta_{n,\theta}$ *converge en loi vers une variable aléatoire* η_θ *uniformément en* θ si pour toute fonction φ continue et bornée

$$\sup_\theta |\mathbf{E}_\theta \varphi(\eta_{n,\theta}) - \mathbf{E}\varphi(\eta_\theta)| \to 0 \qquad (1)$$

lorsque $n \to \infty$.

Nous écrirons cette relation sous la forme « $\eta_{n,\theta} \Rightarrow \eta_\theta$ *uniformément en* θ ». Nous attribuerons la même signification à la relation « $\eta_{n,\theta} \in \mathbf{G}_\theta$ uniformément en θ », où $\mathbf{G}_\theta$ est la distribution de η_θ.

Nous laissons au lecteur le soin de s'assurer que si les fonctions de répartition de η_θ sont continues uniformément par rapport à θ, la relation (1) est équivalente à

$$\sup_{\theta,x} |\mathbf{P}_\theta(\eta_{n,\theta} < x) - \mathbf{P}(\eta_\theta < x)| \to 0.$$

Signalons qu'il y a équivalence entre la convergence uniforme $\eta_{n,\theta} \underset{P_\theta}{\to} a(\theta)$ et la convergence uniforme en loi $\eta_{n,\theta} \Rightarrow a(\theta)$, où $a(\theta)$ est une variable aléatoire dégénérée.

Signalons encore que la convergence uniforme est justiciable des théorèmes de continuité fondamentaux. Si par exemple H est une fonction continue, la convergence uniforme $\eta_{n,\theta} \Rightarrow \eta_\theta$ entraîne la convergence uniforme

$$H(\eta_{n,\theta}) \Rightarrow H(\eta_\theta). \tag{2}$$

Ces assertions découlent directement des définitions.

L'annexe V contient les démonstrations des théorèmes limites « uniformes » suivants.

Soit $X \in \mathbf{P}_\theta$ et soit $a(x, \theta)\colon \mathscr{X} \times \Theta \to R^l$ une fonction vectorielle mesurable donnée. Considérons les sommes

$$s_n(\theta) = \sum a(\mathrm{x}_i, \theta)$$

de vecteurs aléatoires indépendants qui sont fonctions du paramètre $\theta \in \Theta$ soit directement par l'intermédiaire de $a(x, \theta)$, soit par l'intermédiaire de la distribution $\mathbf{P}_\theta$ de x_i.

Rappelons que l'intégrale $\int \psi(x, \theta)\mathbf{P}_\theta(dx)$ est dite convergente uniformément en θ dans le domaine Θ si

$$\sup_{\theta\in\Theta} \int_{|\psi(x,\theta)| > N} |\psi(x, \theta)|\mathbf{P}_\theta(dx) \to 0$$

lorsque $N \to \infty$.

THÉORÈME 1 (loi uniforme des grands nombres). *Si l'intégrale* $a(\theta) = \int a(x, \theta)\mathbf{P}_\theta(dx)$ *converge uniformément en* $\theta \in \Theta$, *alors*

$$\frac{s_n(\theta)}{n} \underset{P_\theta}{\to} a(\theta)$$

uniformément en θ *lorsque* $n \to \infty$.

COROLLAIRE 1. *Si la suite* $\{\theta_n\} \subset \Theta$ *et si l'on se place dans les conditions du théorème* 1, *alors*

$$\mathbf{P}_{\theta_n}\left(\left|\frac{s_n(\theta_n)}{n} - a(\theta_n)\right| > \epsilon\right) \to 0.$$

On notera ce fait par

$$\frac{s_n(\theta_n)}{n} - a(\theta_n) \underset{P_{\theta_n}}{\to} 0.$$

En étudiant le théorème limite central pour les sommes $s_n(\theta)$, on aura intérêt à admettre que $a(\theta) = 0$. (Ceci n'est pas une restriction de la généralité, puisque nous pouvons envisager de nouveaux termes $a^1(\mathrm{x}_i, \theta) = a(\mathrm{x}_i, \theta) - a(\theta)$.) Posons $\sigma^2(\theta) = \mathbf{E}_\theta(a^T(\mathrm{x}_1, \theta)\, a(\mathrm{x}_1, \theta))$ et désignons par $a_j(\mathrm{x}_1, \theta)$, $j = 1, 2, \ldots, l$, les coordonnées des vecteurs $a(\mathrm{x}_1, \theta)$.

THÉORÈME 2 (théorème limite central uniforme). *Si les intégrales* $\int a_j^2(x, \theta)\mathbf{P}_\theta(dx)$, $j = 1, \ldots, l$, *convergent uniformément dans* Θ, *alors*

$$\eta_{n,\theta} = \frac{s_n(\theta)}{\sqrt{n}} \Rightarrow \eta_\theta \in \Phi_{0,\sigma^2(\theta)}$$

uniformément en θ.

2. Variantes uniformes des théorèmes sur les propriétés asymptotiques du rapport de vraisemblance et les estimateurs du maximum de vraisemblance. Remarquons préalablement que si les conditions (RR) sont remplies, les résultats du § 23 sont uniformes en θ, puisque les seconds membres des inégalités des théorèmes 23.1 à 23.3 (et des théorèmes 28.1 à 28.3) sont indépendants de θ.

Passons aux résultats des §§ 24, 28 relatifs au comportement asymptotique de $Z(u/\sqrt{n})$.

Les propositions des lemmes 24.1, 28.1, 24.2, 28.2 peuvent être rendues uniformes en θ.

LEMME 1. *Lorsque* $\Delta \to 0$

$$\sup_\theta \mathbf{E}_\theta \omega_\Delta''(\mathrm{x}_1) \to 0, \tag{3}$$

où $\omega_\Delta''(\mathrm{x}_1)$ *est le plus grand module de continuité des fonctions* $l_{ij}''(x, \theta)$.

DÉMONSTRATION. La relation (3) à θ fixe a été prouvée dans le lemme 28.1. Si l'on admet la non-uniformité en θ, on peut exhiber un $\varepsilon > 0$ et des

suites $\theta_n \to \theta \in \Theta$, $\Delta_n \to 0$, tels que

$$E_{\theta_n}\omega''_{\Delta_n}(x_1) > \epsilon. \tag{4}$$

En posant pour simplifier $\omega''_{\Delta_n}(x_1) = \omega''$, on obtient

$$E_{\theta_n}\omega'' = E_{\theta_n}(\omega''; f_{\theta_n}(x_1) \leqslant 2f_\theta(x_1)) + E_{\theta_n}(\omega''; f_{\theta_n}(x_1) >$$
$$> 2f_\theta(x_1), l(x_1) \leqslant N) + E_{\theta_n}(\omega''; f_{\theta_n}(x_1) > 2f_\theta(x_1), l(x_1) > N).$$

Le premier terme est $\leqslant 2E_\theta\omega''$ et converge vers 0 en vertu du lemme 28.1. Le second est $\leqslant 2NJ_n$, où

$$J_n = \int\limits_{f_{\theta_n}(x) > 2f_\theta(x)} f_{\theta_n}(x)\mu(dx) = 1 - \int\limits_{f_{\theta_n}(x) \leqslant 2f_\theta(x)} f_{\theta_n}(x)\mu(dx) \to 0$$

d'après le théorème de la convergence dominée. Le dernier terme enfin est $\leqslant E_{\theta_n}(2l(x_1); l(x_1) > N)$ et en vertu des conditions (RR) peut être rendu aussi petit que l'on veut moyennant un choix convenable de N. Cette contradiction avec (4) prouve le lemme.

LEMME 2. *Le lemme* 28.2 *reste en vigueur si l'on remplace la convergence presque sûre par la convergence* $\gamma_n(\delta_n, \theta) \underset{P_\theta}{\rightrightarrows} 0$, $\gamma_n(\delta_n, \hat\theta^*) \underset{P_\theta}{\rightrightarrows} 0$ *uniforme en* θ.

DÉMONSTRATION. On suivra la démonstration du lemme 28.2. Remarquons préalablement qu'en vertu du théorème 1 et de la convergence uniforme de l'intégrale dans les conditions (RR),

$$L''(X, \theta)/n \underset{P_\theta}{\to} -I(\theta)$$

uniformément en θ (la convergence des matrices porte sur les éléments). Par ailleurs, des théorèmes 23.3 et 28.3 il découle que $\hat\theta^* \underset{P_\theta}{\rightrightarrows} \theta$ uniformément en θ. De là il résulte que dans la relation $\gamma_n(\delta_n, \theta) \underset{P_\theta}{\to} 0$ (cf. lemme 28.2) on peut remplacer $I(\theta)$ par $L''(\theta)/n$ et par $I(\hat\theta^*)$.

En vertu de l'inégalité (28.7), l'estimation de $\gamma_n(\delta_n, \theta)$ se ramène à celle de

$$\bar\omega''_{\delta_n}(X) = \frac{1}{n}\sum_{i=1}^{n}\omega''_{\delta_n}(x_i, \theta),$$

où $\omega''_\Delta(x, \theta)$ est le plus grand module de continuité des fonctions $l''_{ij}(x, \theta)$. L'inégalité de Tchébychev nous donne

$$\sup_\theta P_\theta(\bar\omega''_{\delta_n}(X) > \epsilon) \leqslant \frac{1}{\epsilon}\sup_\theta E_\theta\omega''_{\delta_n}(x_1, \theta).$$

Mais le lemme 1 nous dit que $\sup_\theta \mathbf{E}_\theta \omega''_\Delta(x_1, \theta) \to 0$ lorsque $\Delta \to 0$. Ceci prouve que

$$\bar{\omega}''_{\delta_n}(X) \underset{P_\theta}{\rightrightarrows} 0, \quad \gamma_n(\delta_n, \theta) \underset{P_\theta}{\rightrightarrows} 0 \tag{5}$$

uniformément en θ.

Les inégalités (24.6) ramènent l'estimation de $\gamma_n(\delta_n, \hat{\theta}^*)$ à celle de $\bar{\omega}''_{\delta_n + |\hat{\theta}^* - \theta|}(X)$. Comme $\hat{\theta}^* - \theta \underset{P_\theta}{\rightrightarrows} 0$ uniformément en θ, on déduit de (5) que

$$\bar{\omega}''_{\delta_n + |\hat{\theta}^* - \theta|}(X) \underset{P_\theta}{\rightrightarrows} 0, \quad \gamma_n(\delta_n, \hat{\theta}^*) \underset{P_\theta}{\rightrightarrows} 0$$

uniformément en θ. ◄

THÉORÈME 3 (analogue du théorème 28.4). *Si les conditions* (*RR*) *sont remplies, le théorème* 28.4 *reste en vigueur moyennant les changements suivants :* $\epsilon_n(X, \theta) \underset{P_\theta}{\rightrightarrows} 0$ *uniformément en* θ, $\xi_n \Subset \Phi_{0, I(\theta)}$, $2Y(u^*) \Subset \mathbf{H}_k$ *uniformément en* θ.

DÉMONSTRATION. Elle repose entièrement sur le lemme 2 au même titre que la démonstration du théorème 28.4 repose sur le lemme 28.2. On établira donc ce théorème en procédant, dans la démonstration du théorème 28.4, aux changements évidents entraînés par la substitution (résultant du lemme 28.2) de la convergence uniforme $\varepsilon_n(X, \theta) \underset{P_\theta}{\rightrightarrows} 0$ à la convergence $\varepsilon_n(X, \theta) \underset{p.s.}{\rightarrow} 0$. Ajoutons par ailleurs que

$$\xi_n = \frac{1}{n} \sum_{i=1}^{n} l'(x_i, \theta) \Subset \Phi_{0, I(\theta)}$$

uniformément en θ en vertu du théorème 2 et de la convergence uniforme (28.6) de l'intégrale $I(\theta)$ (ceci est la matrice des moments d'ordre deux pour $l'(x_1, \theta)$ qui résulte des conditions (*RR*) (cf. Annexe VI). De là et des remarques relatives à (2), on déduit la convergence uniforme

$$2Y(u^*) \Subset \mathbf{H}_k. \ ◄$$

Les changements effectués dans le théorème 3 (par rapport au théorème 28.4) peuvent être introduits dans les théorèmes 28.5 et 28.6.

Le théorème 3 admet les deux corollaires suivants.

THÉORÈME 4.

$$u^* = \sqrt{n}(\hat{\theta}^* - \theta) \Subset \Phi_{0, I^{-1}(\theta)} \tag{6}$$

uniformément en θ. *Ceci étant, pour toute fonction* $w(x)$ *continue presque partout par rapport à la mesure de Lebesgue et telle que* $|w(x)| < Ce^{\beta|x|^{2/k}}$

(*la valeur* $\beta > 0$ *est définie dans le théorème* 28.2), *on a*

$$\sup_\theta |\mathbf{E}_\theta w(u^*) - \mathbf{E} w(\eta_\theta)| \to 0, \tag{7}$$

où $\eta_\theta \in \Phi_{0, I^{-1}(\theta)}$.

DÉMONSTRATION. La première assertion résulte des relations

$$u^* = \xi_n I^{-1}(\theta)(E + \epsilon_n(X, \theta)),$$

$$|\epsilon_n(X, \theta)| \underset{P_\theta}{\rightrightarrows} 0, \quad \xi_n \in \Phi_{0, I(\theta)},$$

uniformes en θ qui sont contenues dans le théorème 3.

Pour établir la deuxième assertion, on admettra que (7) est mise en défaut. Il existe alors un $\delta > 0$ et une suite $\theta_n \to \theta \in \Theta$ tels que

$$|\mathbf{E}_{\theta_n} w(u^*) - \mathbf{E} w(\eta_{\theta_n})| > \delta \tag{8}$$

pour tous les n. Or $\Phi_{0, I^{-1}(\theta_n)} \Rightarrow \Phi_{0, I^{-1}(\theta)}$ et par suite, la $\mathbf{P}_{\theta_n}$-distribution de u^* (resp. de $w(u^*)$) converge, en vertu de (6), faiblement vers la distribution de η_θ (resp. de $w(\eta_\theta)$). Par ailleurs, le corollaire 23.2 (cf. également § 28) nous dit que

$$\sup_\theta \mathbf{E}_\theta w^{3/2}(u^*) \leqslant \sup_\theta \mathbf{E}_\theta \exp\{3(u^*)^2\beta/4\} < c_1 < \infty.$$

De là et des théorèmes de continuité des moments, il vient

$$\mathbf{E}_{\theta_n} w(u^*) \to \mathbf{E} w(\eta_\theta).$$

Cette relation contredit (8), puisque $\mathbf{E} w(\eta_{\theta_n}) \to \mathbf{E} w(\eta_\theta)$. ◄

Supposons que $A_n \subset \mathscr{X}^n$.

THÉORÈME 5. *Si* $\mathbf{P}_\theta(A_n) \to 0$, *pour tout* N *fixe on a*

$$\sup_{|u| \leqslant N} \mathbf{P}_{\theta + u/\sqrt{n}}(A_n) \to 0.$$

Cette propriété des suites de distributions $\mathbf{P}_{\theta + u/\sqrt{n}}$, $n \to \infty$, est appelée *propriété de contingence* (cf. [71]). Nous l'utiliserons au chapitre 3.

DÉMONSTRATION. On a

$$\mathbf{P}_{\theta + u/\sqrt{n}}(A_n) = \mathbf{E}_\theta\{Z(u/\sqrt{n}); A_n\} \leqslant$$
$$\leqslant \mathbf{E}_\theta(Z(u/\sqrt{n}); A_n \cap \{Y(u) \leqslant c\}) + \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c) \leqslant$$
$$\leqslant e^c \mathbf{P}_\theta(A_n) + \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c).$$

Puisque $\mathbf{P}_\theta(A_n) \to 0$, pour prouver ce théorème il faut étudier seulement $\sup_{|u| \leqslant N} \mathbf{P}_{\theta + u/\sqrt{n}}(Y(u) > c)$. D'après le théorème 3, on a uniformément en u

$$Y(u) := (\xi_n, u) - \frac{1}{2} u I(\theta) u^T (1 + \epsilon_{in}(X, \theta + u/\sqrt{n})) \in \Phi_{-\frac{1}{2}\sigma^2, \sigma^2}, \tag{9}$$

où $\sigma^2 = uI(\theta)u^T \leqslant N^2\Lambda_k(\theta)$ pour $|u| \leqslant N$ et $\Lambda_k(\theta)$ est la plus grande valeur propre de la matrice $I(\theta)$. Vu que

$$\Phi_{-\frac{1}{2}\sigma^2,\sigma^2}(]c,\infty[) \leqslant \Phi_{0,\sigma^2}(]c,\infty[),$$

il vient en vertu de la convergence uniforme dans (9)

$$\lim_{n\to\infty} \sup_{|u|\leqslant N} \mathbf{P}_{\theta+u/\sqrt{n}}(Y(u) > c) \leqslant \sup_{|u|\leqslant N} \Phi_{0,\sigma^2}(]c,\infty[) = \Phi_{0,N^2\Lambda_k(\theta)}(]c,\infty[).$$

Cette valeur peut être rendue aussi petite que l'on veut moyennant un choix convenable de c. ◄

3. Quelques corollaires.

1) Le théorème 25.3 affirme en particulier que $\hat{\theta}^* \in \tilde{K}^\circ$, où $\tilde{K}^\circ$ est la classe des estimateurs asymptotiquement centraux, définie par la relation (on étudie le cas scalaire)

$$\mathbf{P}_\theta(\hat{\theta}^* > \theta) \to 1/2$$

uniformément en θ. Du théorème 4 il résulte que cette partie du théorème 25.3 est valable, puisque

$$\mathbf{P}_\theta(\hat{\theta}^* > \theta) = \mathbf{P}_\theta(\sqrt{n}(\hat{\theta}^* - \theta)I^{-1/2}(\theta) > 0) \to \Phi_{0,1}(]0,\infty[) = 1/2$$

uniformément en θ. ◄

2) Au § 25 nous avons formulé le théorème 7 de minimaxialité asymptotique de $\hat{\theta}^*$. Pour prouver ce théorème il reste à établir le lemme 25.1 qui dit que

$$\lim_{n\to\infty} . \sup_{\theta\in\Gamma} \mathbf{E}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} I^{-1}(\theta), \tag{10}$$

où Γ est un intervalle fermé quelconque de Θ. Mais cette proposition résulte directement de la convergence $\mathbf{E}_\theta n(\hat{\theta}^* - \theta)^2 \to I^{-1}(\theta)$ uniforme en $\theta \in \Theta$ qui rend licite le passage à la limite sous le signe $\sup\limits_{\theta\in\Gamma}$:

$$\lim_{n\to\infty} . \sup_{\theta\in\Gamma} \mathbf{E}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} . \lim_{n\to\infty} \mathbf{E}_\theta n(\hat{\theta}^* - \theta)^2 = \sup_{\theta\in\Gamma} I^{-1}(\theta). \quad ◄$$

Nous avons une proposition identique à (10) assurant la minimaxialité asymptotique de $\hat{\theta}^*$ dans le cas multidimensionnel :

$$\lim_{n\to\infty} . \sup_{\theta\in\Gamma} \mathbf{E}_\theta n(\hat{\theta}^* - \theta)V(\hat{\theta}^* - \theta)^T = \sup_{\theta\in\Gamma} \sum v_{ij} I_{ij}^{-1}(\theta),$$

$$\|I_{ij}^{-1}(\theta)\| = I^{-1}(\theta),$$

pour toute matrice V.

§ 30*. Sur les problèmes de statistique relatifs aux échantillons de taille aléatoire. Estimation séquentielle

L'exemple 18.3 montre que les échantillons de taille aléatoire se présentent souvent en pratique et sont naturels. Un autre exemple est lié à l'estimation séquentielle. Ce genre d'estimation est utilisé dans les cas où les observations sont séquentielles et qu'il faut minimiser leur nombre en raison, par exemple, de leur coût élevé. Ceci étant, la règle d'estimation (c'est-à-dire la construction d'un estimateur θ^*) doit être définie en même temps que la règle d'arrêt des observations. Ces règles peuvent être de nature différente: on peut par exemple sommer des coûts $c(x_i)$ donnés des observations x_i tant qu'on n'a pas atteint une quantité donnée t. Dans ce cas, la date ν d'arrêt (le numéro de la dernière observation ou la taille de l'échantillon) sera définie comme suit :

$$\nu = \min \left\{ k : \sum_{i=1}^{k} c(x_i) \geqslant t \right\},$$

cette quantité désigne la « date du premier passage du niveau t » dans une promenade de sauts $c(x_i)$ (cf. [11], chap. 8). On peut sommer l' « information » $I(x_i, \theta) = (l'(x_i, \theta))^2$ et cesser les observations lorsqu'on aura atteint un niveau donné, et ainsi de suite.

Dans ces exemples, ν est un instant markovien, c'est-à-dire $\{\nu > n\} \in \in \sigma(x_1, \ldots, x_n)$. Ceci est l'une des principales conditions que l'on pose en étudiant les problèmes d'estimation séquentielle. Si cette condition est réalisée avec d'autres moins fondamentales, l'inégalité de Rao-Cramer reste en vigueur sous la forme

$$\mathbf{V}_\theta \theta^* \geqslant \frac{1}{I(\theta)\mathbf{E}\nu},$$

où $\theta^* = \theta^*(x_1, \ldots, x_\nu)$ est un estimateur sans biais de θ, $I(\theta)$ la quantité d'information de Fisher. La démonstration de cette inégalité est identique à celles du § 16; il faut seulement se servir de l'identité de Wald (cf. [11]) pour calculer la quantité d'information de Fisher contenue dans l'échantillon $(x_1, \ldots, x_\nu)$.

Si ν dépend d'un paramètre t comme dans l'exemple 18.3, de sorte que $\nu \underset{\text{p.s.}}{\to} \infty$ lorsque $t \to \infty$, il est alors possible de construire des estimateurs asymptotiquement optimaux dont l'erreur quadratique moyenne est asymptotiquement équivalente à $(I(\theta)\mathbf{E}\nu)^{-1}$.

§ 31. Estimation par intervalle

1. Définitions. Jusqu'ici nous avons étudié les propriétés et les procédés de détermination des meilleurs estimateurs *ponctuels* du paramètre inconnu θ qui définit dans une famille $\mathscr{P} = \{\mathbf{P}_\theta\}$ une distribution $\mathbf{P}_\theta$ associée à l'échantillon X. Les estimations ponctuelles sont utilisées dans les cas où il faut désigner un nombre θ^* appelé à remplacer le paramètre inconnu θ.

Il existe une autre approche assez répandue de ce problème.

On admettra que θ est un paramètre scalaire (le cas vectoriel sera examiné au n° 6). On sait qu'il est impossible de déterminer exactement θ au vu de l'échantillon donné. Mais on pourrait tenter d'indiquer un intervalle $]\theta^-, \theta^+[$ qui contiendrait la valeur inconnue de θ avec une probabilité assez élevée donnée *a priori*. Il est évident qu'on aura intérêt à ce que cet intervalle soit le plus étroit possible. Dans de nombreux problèmes on demande, par exemple, en augmentant la taille de l'échantillon, de construire un intervalle $]\theta^-, \theta^+[$ dont la longueur soit au plus égale à une quantité donnée.

DÉFINITION 1. Supposons que pour un $\epsilon > 0$ il existe des variables aléatoires $\theta^\pm = \theta^\pm(\epsilon, X)$ telles que

$$\mathbf{P}_\theta(\theta^-(\epsilon, X) < \theta, \theta^+(\epsilon, X) > \theta) \geqslant 1 - \epsilon. \qquad (1)$$

L'intervalle $]\theta^-, \theta^+[$ s'appelle alors *intervalle de confiance au seuil* $1 -$ $- \epsilon$ *pour l'estimation de* θ.

Il est évident que (1) peut être mise sous la forme

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) \geqslant 1 - \epsilon.$$

L'événement contenu sous le signe de la probabilité consiste en ce que *l'intervalle aléatoire* $]\theta^-, \theta^+[$ *recouvre la valeur inconnue de* θ. Il serait moins correct de lire cet événement « θ tombe dans l'intervalle $]\theta^-, \theta^+[$ », puisque θ n'est pas aléatoire.

Les valeurs $\theta^\pm$ s'appellent *bornes de l'intervalle de confiance*, le nombre $1 - \epsilon$, *niveau* ou *seuil de confiance*.

Ainsi l'estimation par intervalle diffère de l'estimation ponctuelle sur les deux points suivants :

1) L'estimation par intervalle est moins « exacte », car elle indique tout un ensemble de valeurs éventuelles de θ.

2) L'affirmation « $\theta \in]\theta^-, \theta^+[$ avec une probabilité $\geqslant 1 - \epsilon$ » est vraie, tandis que l'événement $\{\theta = \theta^*\}$ est en général de probabilité nulle.

Pour ϵ on prend généralement un petit nombre. On construit $\theta^\pm(\epsilon, X)$, puis on décrète au vu de l'échantillon que $\theta \in]\theta^-(\epsilon, X), \theta^+(\epsilon, X)[$. En procédant ainsi on se trompera au cours de nombreuses répétitions de l'expérience environ dans $100\,\epsilon$ % des cas. Si par exemple $\epsilon = 0{,}001$, l'erreur se produira environ une fois sur mille cas.

En décrétant que la relation $\theta \in]\theta^-, \theta^+[$ est vraie, on se sert du fait que si un événement est de probabilité ε et que ε soit petit, il est pratiquement impossible que cet événement se produise en une seule épreuve. Le passager qui prend place dans un avion en est fermement convaincu. Il lui suffit de savoir que la probabilité que le vol se termine normalement soit élevée (il sait en effet que cette probabilité n'est pas égale à 1). Cette approche repose justement à la base de nombreuses procédures statistiques.

Nous commencerons par mettre en évidence le cas où la construction des intervalles de confiance est naturelle et n'apporte aucune complication. Nous avons en vue le cas bayésien qui a déjà été envisagé dans les §§ 10, 11 et 20.

2. Construction des intervalles de confiance dans le cas bayésien. On admettra que le paramètre θ est *aléatoire* et de densité *a priori* $q(t)$ par rapport à une mesure λ sur Θ. On demande de construire un intervalle de confiance pour la valeur retenue de θ au vu d'un échantillon $X \in \mathbf{P}_\theta$.

Si la condition (A_μ) est remplie, on sait du § 10 qu'il existe alors une distribution *a posteriori* de θ (conditionnelle par rapport à X) de densité

$$q(t|X) = \frac{f_t(X)q(t)}{\int f_u(X)q(u)\lambda(du)}$$

par rapport à la mesure λ. Ceci exprime que pour $\theta^\pm(\varepsilon, X)$ il suffit de prendre deux nombres quelconques $\theta^\pm$ pour lesquels

$$\int_{\theta^-}^{\theta^+} q(u|X)\lambda(du) = 1 - \varepsilon$$

(ou $\geqslant 1 - \varepsilon$ si $\int_{-\infty}^{t} q(u|X)\lambda(du)$ varie de façon discrète en fonction de t). En d'autres termes, pour θ^- et θ^+ il faut prendre les quantiles de la distribution *a posteriori* respectivement d'ordre $1 - \varepsilon_2$ et ε_1 pour des ε_1 et ε_2 tels que $\varepsilon_1 + \varepsilon_2 = \varepsilon$.

Contrairement au cas non bayésien, dans la relation $\theta^- \leqslant \theta \leqslant \theta^+$ les trois éléments sont aléatoires : les bornes $\theta^\pm$ et la quantité θ elle-même.

Il est immédiat de voir que la procédure décrite donne lieu à un certain arbitraire dans le choix des nombres ε_1 et ε_2. Cet arbitraire est parfois écarté par la position même du problème, par exemple lorsqu'il faut déterminer seulement la borne supérieure ou inférieure de l'intervalle de confiance. Dans ce cas, il faut poser ε_1 ou ε_2 égal à 0 et rendre infinie la borne correspondante. Si les bornes sont symétriques, il faut choisir ε_i de façon à rendre l'intervalle $]\theta^-, \theta^+[$ le plus petit possible. Pour les distributions $q(t|X)$ proches de distributions symétriques, ceci a lieu pour $\varepsilon_1 = \varepsilon_2 = \varepsilon/2$.

3. Construction des intervalles de confiance dans le cas général. Intervalles de confiance asymptotiques. Les principales méthodes de construction des intervalles de confiance utilisent les estimateurs ponctuels. Nous commencerons par étudier l'approche asymptotique de construction de ces intervalles.

DÉFINITION 2. Soit $X = [X_\infty]_n \in \mathbf{P}_\theta$ et supposons que pour $\epsilon > 0$ il existe des variables aléatoires $\theta^{\pm}(\epsilon, X)$ telles que

$$\lim_{n\to\infty}\inf \mathbf{P}_\theta(\theta^-(\epsilon, X) < \theta < \theta^+(\epsilon, X)) \geqslant 1 - \epsilon . \tag{2}$$

L'intervalle $]\theta^-, \theta^+[$ s'appelle alors *intervalle de confiance asymptotique au seuil* $1 - \epsilon$.

Dans cette définition il est nécessaire de souligner qu'il est question en fait d'une suite d'intervalles $]\theta_n^-, \theta_n^+[$ définis pour chaque n. Formellement, la notion d'intervalle de confiance asymptotique appliquée à un échantillon de taille fixée est peu intéressante. Il n'empêche qu'on se sert de la relation (2) pour les grands n, au même titre que du théorème limite central pour le calcul approché des distributions des sommes d'un nombre fini de variables aléatoires.

Nous avons vu dans les paragraphes précédents que la plupart des estimateurs ponctuels étudiés étaient asymptotiquement normaux. Nous construirons plus bas des intervalles de confiance asymptotiques basés sur ces estimateurs.

Soit θ^* un estimateur asymptotiquement normal :

$$(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2(\theta)}, \tag{3}$$

et soit $\sigma(\theta)$ une fonction continue. Comme $\theta^* \underset{p}{\to} \theta$, la dernière condition exprime que $\sigma(\theta^*) \underset{p}{\to} \sigma(\theta)$. De là et de la relation (3) il s'ensuit en vertu du deuxième théorème de continuité que

$$\frac{(\theta^* - \theta)\sqrt{n}}{\sigma(\theta^*)} \in \Phi_{0,1}. \tag{4}$$

Désignons par λ_δ le quantile d'ordre $1 - \delta$ de la distribution normale, c'est-à-dire le nombre tel que $\Phi_{0,1}(]-\infty, \lambda_\delta[) = 1 - \delta$, ou $\mathbf{P}(|\xi| < \lambda_\delta) = 1 - 2\delta$ si $\xi \in \Phi_{0,1}$. Désignons provisoirement pour abréger $\lambda_{\epsilon/2}$ par β, où $\epsilon > 0$ est fixe et donné. De (4) il vient alors

$$\lim_{n\to\infty} \mathbf{P}_\theta\left(\left|\frac{(\theta^* - \theta)\sqrt{n}}{\sigma(\theta^*)}\right| < \beta\right) = 1 - \epsilon .$$

Or cette relation peut être mise sous la forme

$$\lim_{n\to\infty} \mathbf{P}_\theta(\theta^* - \beta\sigma(\theta^*)/\sqrt{n} < \theta < \theta^* + \beta\sigma(\theta^*)/\sqrt{n}) = 1 - \epsilon .$$

Donc, les nombres

$$\theta^{\pm} = \theta^* \pm \beta\sigma(\theta^*)/\sqrt{n} \quad (5)$$

vérifient la définition 2 et par suite sont les bornes d'un intervalle de confiance asymptotique au seuil $1 - \epsilon$.

Si maintenant nous construisons l'intervalle (5) pour un échantillon fixe X de taille n, son seuil sera différent de ε, mais cette différence sera petite si n est assez grand. Il faut donc manipuler les intervalles de confiance asymptotiques avec une certaine prudence en s'assurant préalablement à partir de quelles valeurs de n la probabilité de l'événement $\{\theta \in]\theta^-, \theta^+[\}$ est suffisamment bien approchée par la valeur limite. En général, plus ε est petit, plus les conditions imposées à la taille n de l'échantillon sont strictes. La taille nécessaire dépend aussi de la distribution $\mathbf{P}_\theta$ et de la statistique θ^*.

EXEMPLE 1. Supposons que $X \in \mathbf{\Gamma}_{\alpha,1}$ et utilisons l'estimateur efficace $\alpha^* = \dfrac{n-1}{n\bar{x}}$. Dans les exemples 4.1 et 16.1, on a établi que

$$\mathbf{E}_\alpha\alpha^* = \alpha, \quad \mathbf{V}_\alpha\alpha^* = \frac{\alpha^2}{n-2},$$

de sorte que $\sigma^2(\alpha) = \alpha^2$. La relation (5) nous donne

$$\alpha^{\pm} = \frac{n-1}{n\bar{x}}(1 \pm \beta/\sqrt{n}). \quad (6)$$

A quoi est égal le seuil de cet intervalle?

Il nous faut trouver la probabilité $\mathbf{\Gamma}_{\alpha,1}$ de la double inégalité

$$\frac{n-1}{n\bar{x}}(1 - \beta/\sqrt{n}) < \alpha < \frac{n-1}{n\bar{x}}(1 + \beta/\sqrt{n})$$

où, ce qui est équivalent, de la double inégalité

$$1 - \beta/\sqrt{n} < \frac{n\alpha\bar{x}}{n-1} < 1 + \beta/\sqrt{n},$$

où $n\alpha\bar{x} \in \mathbf{\Gamma}_{1,n}$. Le paramètre α étant paramètre d'échelle, il vient $2n\alpha\bar{x} \in \in \mathbf{\Gamma}_{1/2,n} = \mathbf{H}_{2n}$. Donc, le seuil exact de l'intervalle (6) est égal à

$$\int_{2(n-1)(1-\beta/\sqrt{n})}^{2(n-1)(1+\beta/\sqrt{n})} \gamma_{1/2,n}(x)dx, \quad (7)$$

où $\gamma_{1/2,n}$ est défini dans le § 2 *).

---

*) La remarque que $\mathbf{\Gamma}_{1/2,n} = \mathbf{H}_{2n}$ est utile, car elle permet d'appliquer au calcul de $\mathbf{\Gamma}_{\alpha,\lambda}$ (si 2λ est entier) les tables de la distribution χ^2 citées en annexe et dans de nombreux aide-mémoire de statistique mathématique.

Pour $\epsilon = 0{,}05$ et $n = 30$, on a $\beta = 1{,}96$, $(n - 1)(1 - \beta/\sqrt{n})/n \approx 0{,}6201$, $(n - 1)(1 + \beta/\sqrt{n})/n \approx 1{,}3126$.

Donc, l'intervalle de confiance asymptotique au seuil $1 - \epsilon = 0{,}95$ pour $n = 30$ est l'intervalle $]0{,}620/\bar{x}, 1{,}313/\bar{x}[$.

Si l'on utilise les tables de la distribution du χ^2 à 60 degrés de liberté, on trouve en vertu de (7) que le seuil exact de cet intervalle de confiance est égal (au millième près) à $0{,}937 = 1 - 0{,}063$. Ceci étant, les « contributions » des bornes de gauche et de droite de l'intervalle de confiance ne sont pas égales (comparer avec l'approximation normale) et valent respectivement 0,010 et 0,053.

Pour $n = 50$ l'intervalle de confiance asymptotique au seuil 0,95 sera $]0{,}708/\bar{x}, 1{,}252/\bar{x}[$. Son seuil exact sera égal à $0{,}942 = 1 - 0{,}058$ (les « contributions » seront égales respectivement à 0,014 et 0,044). Il est clair que si l'on continue de faire croître n, les « contributions » se rapprocheront de 0,025.

Revenons à l'intervalle de confiance (5) construit à l'aide de l'estimateur asymptotiquement normal θ^*. Contrairement au cas bayésien l'arbitraire est introduit ici par le choix de l'estimateur θ^*. La forme des bornes de l'intervalle montre qu'on peut obtenir un intervalle de dimensions voulues soit en faisant croître la taille de l'échantillon n (ce qui n'est pas toujours possible), soit en réduisant $\sigma(\theta^*)$. On est ainsi conduit à l'importante conclusion suivante : à tailles égales le meilleur intervalle de confiance sera fourni par l'estimateur dont la variance $\sigma(\theta)$ est la plus petite. Donc, *les meilleurs intervalles de confiance asymptotiques seront donnés par les estimateurs asymptotiquement efficaces.*

Si les conditions (RR) sont remplies et θ^* appartient à la classe $\tilde{K}_0 \cap \cap K_{\Phi,2}$ (cf. §§ 8, 16), les bornes du meilleur intervalle de confiance asymptotique sont

$$\theta^{\pm} = \theta^* \pm \beta/\sqrt{nI(\theta^*)}\,,$$

où θ^* est une estimation asymptotiquement efficace quelconque, par exemple une estimation par le maximum de vraisemblance.

D'autres méthodes de construction des intervalles de confiance asymptotiques seront envisagées au n° 6.

4. Construction d'un intervalle de confiance exact à l'aide d'une statistique donnée. Supposons que pour statistique nous avons choisi un estimateur θ^*. Il est naturel de chercher un intervalle de confiance symétrique au seuil $1 - \epsilon$, sous la forme $\theta^* \pm \Delta(\epsilon, X)$ ou $\theta^*(1 \pm \Delta(\epsilon, X))$ comme nous l'avons fait dans l'exemple envisagé plus haut. Mais la réalisation de ce plan soulève de grosses difficultés, car dans le cas général les bornes $\pm\Delta(\epsilon, X)$ dépendront du paramètre inconnu θ : en effet $\Delta(\epsilon, X)$ doit être déterminé

à partir de la condition

$$\mathbf{P}_\theta(\theta^* - \Delta(\epsilon, X) < \theta < \theta^* + \Delta(\epsilon, X)) \geqslant 1 - \epsilon$$

dans laquelle θ figure de façon essentielle et assez compliquée, notamment par l'intermédiaire de la distribution $\mathbf{P}_\theta$.

Il faut un procédé spécial pour construire les intervalles de confiance à l'aide de l'estimateur θ^*.

Le procédé proposé plus bas fait intervenir l'estimateur θ^* et une statistique quelconque S. Désignons la distribution de S par $\mathbf{G}_\theta$ et posons $G_\theta(x) = \mathbf{G}_\theta(]-\infty, x[)$.

DÉFINITION 3. On dira qu'une statistique S *dépend en loi monotonement de* θ si pour tous x, $\theta_1 < \theta_2$, on a

$$\mathbf{G}_{\theta_1}(]x, \infty[) \leqslant \mathbf{G}_{\theta_2}(]x, \infty[),$$

ou ce qui est équivalent

$$G_{\theta_1}(x) \geqslant G_{\theta_2}(x). \tag{8}$$

Tous les estimateurs raisonnables θ^* jouissent de cette propriété.

Si la dépendance monotone de $G_\theta(x)$ par rapport à θ est de plus continue, l'équation

$$G_\theta(x) = \gamma$$

admet toujours une solution θ pour tout $\gamma \in]0, 1[$, que nous désignerons par $b(x, \gamma)$.

THÉORÈME 1. *Si* $\epsilon_1 + \epsilon_2 = \epsilon$, *la statistique S dépend monotonement en loi de θ et la fonction $G_\theta(x)$ est continue par rapport à θ et x, alors les valeurs*

$$\theta^- = b(S, 1 - \epsilon_2), \quad \theta^+ = b(S, \epsilon_1)$$

sont les bornes d'un intervalle de confiance au seuil $1 - \epsilon$.

DÉMONSTRATION. Elle est presque évidente. Utilisons le fait que si la fonction de répartition $F(x)$ est continue et $\xi \in F$, alors $F(\xi) \in \mathbf{U}_{0,1}$ $(\mathbf{P}(F(\xi) < x) = \mathbf{P}(\xi < F^{-1}(x)) = F(F^{-1}(x)) = x)$. En vertu de cette remarque, $G_\theta(S) \in \mathbf{U}_{0,1}$ et par suite

$$\mathbf{P}_\theta(\epsilon_1 < G_\theta(S) < 1 - \epsilon_2) = 1 - \epsilon,$$

$$\mathbf{P}_\theta(b(S, 1 - \epsilon_2) < \theta < b(S, \epsilon_1)) = 1 - \epsilon. \quad ◄$$

Il est souvent commode d'« inverser » en deux étapes la fonction $G_\theta(S)$ intervenant dans le théorème. D'abord par rapport à x, c'est-à-dire qu'on définit les quantiles $G_\theta^{-1}(\gamma)$ comme les solutions des équations $G_\theta(x) = \gamma$,

et ensuite on résout les équations

$$G_\theta^{-1}(\epsilon_1) = S, \quad G_\theta^{-1}(1 - \epsilon_2) = S$$

par rapport à θ. Ces équations admettent toujours des solutions, puisque $G_\theta^{-1}(\gamma)$ est monotone et dépend continûment de θ par hypothèse.

La figure 3 représente les courbes $y = G_\theta^{-1}(\epsilon_1)$ et $y = G_\theta^{-1}(1 - \epsilon_2)$ qui définissent pour chaque θ un domaine de valeurs de y auquel la probabilité d'accès est égale à $1 - \epsilon$ pour un estimateur $S = \theta^*$. Comme déjà signalé,

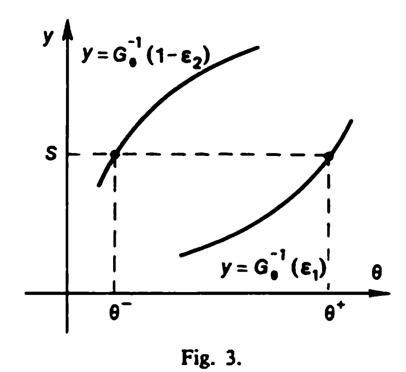


Fig. 3.

la procédure de construction de l'intervalle de confiance se traduit par l'inversion des fonctions

$$y = G_\theta^{-1}(\epsilon_1), \quad y = G_\theta^{-1}(1 - \epsilon_1),$$

c'est-à-dire par la recherche des points d'intersection des courbes représentatives de ces fonctions avec la ligne de niveau $y = S$. Les points d'intersection obtenus nous fournissent l'intervalle $]\theta^-, \theta^+[$ cherché.

Si la condition de continuité est violée (dans le cas notamment de variables aléatoires S discrètes), la procédure exposée et le théorème restent dans l'ensemble en vigueur à la seule différence que dans la définition des quantiles $G_\theta^{-1}(\gamma)$ il faudra satisfaire l'inégalité $\mathbf{G}_\theta(]G_\theta^{-1}(\epsilon_1), G_\theta^{-1}(1 - \epsilon_2)[) \geqslant 1 - \varepsilon$ au lieu de l'égalité. Le théorème 1 devient alors

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) \geqslant 1 - \epsilon,$$

où $\theta^\pm$ sont solutions des équations $G_\theta^{-1}(\epsilon_1) = S$ et $G_\theta^{-1}(1 - \epsilon_2) = S$. L'intervalle $]\theta^-, \theta^+[$ sera comme précédemment appelé intervalle de confiance au seuil $1 - \epsilon$.

Si l'on construit l'intervalle de confiance $]\theta^-, \theta^+[$ à l'aide de l'estimateur θ^*, on voit sur la figure 3 qu'il sera d'autant plus étroit que le sera l'intervalle $]G_\theta^{-1}(\epsilon_1), G_\theta^{-1}(1 - \epsilon_2)[$ ou, ce qui est équivalent, que la distribution de θ^* sera plus concentrée autour de θ. On est conduit donc au même problème

qu'en théorie de l'estimation ponctuelle, savoir la recherche des meilleurs estimateurs θ^*.

La construction des meilleurs intervalles de confiance sera étudiée en détail dans le § 3.8.

La procédure d'inversion de la fonction de répartition $G_\theta(x)$ est assez épineuse en raison de la forme complexe de $G_\theta(x)$ même pour les familles de distributions citées dans le § 2. Aussi le calcul des bornes des intervalles de confiance est-il essentiellement tabulé. Dans l'exemple suivant qui servira à illustrer la construction d'intervalles de confiance d'après la procédure décrite dans le théorème 1, on utilisera pour simplifier une approximation normale.

EXEMPLE 2. Soit $X \in \mathbf{B}_p$. Pour estimateur de p prenons l'estimateur efficace $p^* = \nu/n$, où ν est le nombre de succès dans n épreuves (ν peut être par exemple le nombre de pièces défectueuses dans un lot de n pièces. On demande l'intervalle de confiance pour le pourcentage p de loups).

On a ($q = 1 - p$)

$$G_p(x) = \mathbf{P}_p(p^* < x) = \mathbf{P}_p\left(\frac{\nu - np}{\sqrt{npq}} < \frac{xn - np}{\sqrt{npq}}\right).$$

D'après le théorème 1 il faut résoudre l'équation

$$G_p(p^*) = \gamma \tag{9}$$

pour les valeurs de γ égales à $\epsilon/2$ et $1 - \epsilon/2$. Pour les grands n le théorème limite central affirme que $G_p(x) \approx \Phi((x - p)n/\sqrt{npq})$, où $\Phi(y) = \Phi_{0,1}(]-\infty, y[)$, et par suite l'équation (9) peut être remplacée par son approximation

$$\Phi((p^* - p)n/\sqrt{npq}) = \gamma, \quad \gamma = \epsilon/2,\ 1 - \epsilon/2,$$

ou, ce qui revient au même, $|(p^* - p)n/\sqrt{npq}| = \lambda_{\epsilon/2} = \beta$,

$$(p^* - p)^2 = \beta^2 p(1 - p)/n.$$

Cette équation pour les bornes $p^{\pm}$ de l'intervalle de confiance est l'équation d'une ellipse allongée pour les grands n le long de la bissectrice $p^* - p = 0$. La résolution de cette équation par rapport à p nous donne

$$p^{\pm} \approx p^* \pm \beta\sqrt{p^*(1 - p^*)/n}.$$

On vérifie immédiatement qu'on aurait obtenu le même résultat en appliquant l'approche asymptotique développée au n° 3.

Si n n'est pas assez grand, il faut calculer $G_p(x)$ à l'aide de la formule exacte

$$G_p(x) = \sum_{k < nx} C_n^k p^k (1 - p)^{n-k},$$

et appliquer ensuite la procédure du théorème 1.

Supposons par exemple que $\nu = 2$ pièces sur $n = 10$ sont défectueuses. Pour $\epsilon = 0{,}05$ les bornes exactes de l'intervalle de confiance sont alors égales à $p^- = 0{,}037$ et $p^+ = 0{,}507$. L'importante dimension de cet intervalle s'explique par la maigre information mise à notre disposition.

Si $n = 100$, $\nu = 20$, on obtient pour $\varepsilon = 0{,}05$

$$p^- = 0{,}137, \quad p^+ = 0{,}277.$$

Ces valeurs ont été empruntées dans des tables spéciales donnant la solution numérique du problème des intervalles de confiance d'un nombre p pour divers n et ν (cf. [8]).

5. Autres méthodes de construction des intervalles de confiance. Dans ce numéro on étudiera quelques généralisations de la procédure de construction des intervalles de confiance proposée plus haut.

THÉORÈME 2. *Supposons qu'il existe sur* $\Theta \times \mathscr{X}^n$ *une fonction* $G(\theta, x)$ *telle que la distribution* $\mathbf{H}(B) = \mathbf{P}_\theta(G(\theta, X) \in B)$ *ne dépende pas de* θ. *Supposons par ailleurs que* $G(\theta, x)$ *est continue et monotone par rapport à* θ *pour tout* x.

Supposons enfin que y^- *et* y^+ *vérifient la relation* $\mathbf{H}(]y^-, y^+[) = 1 - \varepsilon$. *Dans ces conditions les statistiques*

$$\theta^- = G^{-1}(y^-, X), \quad \theta^+ = G^{-1}(y^+, X) \quad \text{si } G(\theta, \cdot)\uparrow,$$

et

$$\theta^- = G^{-1}(y^+, X), \quad \theta^+ = G^{-1}(y^-, X) \quad \text{si } G(\theta, \cdot)\downarrow,$$

sont les bornes d'un intervalle de confiance au seuil $1 - \varepsilon$. *Ici* $G^{-1}(y, X)$ *est la solution de l'équation* $G(\theta, X) = y$.

DÉMONSTRATION. La fonction $G(\theta, x)$ étant monotone (on admet pour fixer les idées que $G(\theta, x)$ est strictement croissante par rapport à θ), l'événement $\{G^{-1}(y^-, X) < \theta < G^{-1}(y^+, X)\}$ est confondu avec l'événement $A = \{y^- < G(\theta, X) < y^+\}$.

Par définition de $\mathbf{H}(\cdot)$ et de $y^\pm$ on a

$$\mathbf{P}_\theta(\theta^- < \theta < \theta^+) = \mathbf{P}_\theta(G^{-1}(y^-, X) < \theta < G^{-1}(y^+, X)) =$$
$$= \mathbf{P}_\theta(A) = \mathbf{H}(]y^-, y^+[) = 1 - \epsilon. \blacktriangleleft$$

REMARQUE 1. Dans le théorème 1, pour $G(\theta, X)$ on a envisagé la fonction $G_\theta(S)$. De plus, $\mathbf{H} = \mathbf{U}_{0,1}$.

REMARQUE 2. On peut considérer l'analogue asymptotique du théorème 2 en admettant l'existence d'une suite de fonctions $G_n(\theta, x)$ continues et monotones par rapport à θ, telles que

$$\mathbf{P}_\theta(G_n(\theta, X) \in B) \to \mathbf{H}(B), \quad n \to \infty,$$

où $\mathbf{H}(\cdot)$ ne dépend pas de θ. On obtient alors une méthode de construction des intervalles de confiance asymptotiques, généralisant la méthode de construction des intervalles de confiance asymptotiques à l'aide des estimateurs asymptotiquement normaux développée au n° 3.

Indiquons un autre procédé de choix de la fonction $G(\theta, x)$ intervenant dans le théorème 2.

THÉORÈME 3. *Supposons que* $F_\theta(x) = \mathbf{P}_\theta(\mathrm{x}_1 < x)$ *et que*

1) $F_\theta(x)$ *est continue par rapport à* x *pour tous les* $\theta \in \Theta$,

2) $F_\theta(x)$ *est continue et monotone par rapport à* θ *pour tout* x *fixe. Alors la fonction*

$$G(\theta, x) = -\sum_{i=1}^{n} \ln(F_\theta(x_i))$$

vérifie les conditions du théorème 2.

Si les nombres $y^\pm$ *sont tels que*

$$\frac{1}{\Gamma(n)} \int_{y^-}^{y^+} x^{n-1} e^{-x} dx = 1 - \varepsilon, \tag{10}$$

alors $\theta^\pm = G^{-1}(y^\pm, X)$ *sont les bornes d'un intervalle de confiance au seuil* $1 - \varepsilon$.

DÉMONSTRATION. Assurons-nous que les conditions du théorème 2 sont remplies. Puisque $F_\theta(\mathrm{x}_i)$ est uniformément distribuée sur $[0, 1]$ d'après la condition 1), il vient $-\ln F_\theta(\mathrm{x}_i) \in \mathbf{\Gamma}_{1,1}$ et $G(\theta, X) \in \mathbf{\Gamma}_{1,n}$. Autrement dit, $\mathbf{P}_\theta(G(\theta, X) \in B) = \mathbf{\Gamma}_{1,n}(B)$, et $\mathbf{H} = \mathbf{\Gamma}_{1,n}$ est indépendante de θ. La monotonie et la continuité de $G(\theta, \mathbf{x})$ pour tout $\mathbf{x}$ résultent de la condition 2). Par ailleurs, en vertu de (10)

$$\mathbf{H}([y^-, y^+]) = \mathbf{\Gamma}_{1,n}([y^-, y^+]) = 1 - \varepsilon. \quad ◄$$

On pourrait indiquer d'autres méthodes de construction des intervalles de confiance. Ceci étant, comme en théorie de l'estimation ponctuelle, il se pose aussitôt la question de savoir lequel des intervalles de confiance, si tant est qu'il en existe plusieurs, est le meilleur. Les diverses approches de cette question seront abordées dans le § 3.8. Mais de l'exposé précédent il ressort que la recherche du meilleur intervalle de confiance présente beaucoup d'affinités avec celle de la meilleure estimation ponctuelle. Il est clair aussi que si l'on construit les intervalles de confiance à l'aide des estimateurs ponctuels, il faudra préférer les intervalles construits à l'aide des meilleurs

L'affinité des problèmes d'optimisation des estimations ponctuelle et par intervalle peut être illustrée sur l'exemple de la proposition suivante.

THÉORÈME 4. *Considérons un intervalle de confiance asymptotique* $]\theta^-, \theta^+[$ *au seuil* $1 - \epsilon$ *et supposons que la variable aléatoire* $\theta^* = (\theta^+ + \theta^-)/2$ *est un estimateur asymptotiquement normal et asymptotiquement central* (cf. n° 2 du § 25) *et que la quantité* $\Delta = (\theta^+ - \theta^-)/2$ *est telle que* $\delta = \liminf_{n\to\infty} \sqrt{n}\Delta$ *ne dépende pas de X. Alors* $\delta \geqslant \beta/\sqrt{I(\theta)}$.

Ceci exprime que la longueur de l'intervalle de confiance $]\theta^-, \theta^+[$ ne peut être sensiblement inférieure à $2\beta/\sqrt{nI(\theta)}$, c'est-à-dire à la longueur de l'intervalle au seuil $1 - \varepsilon$, construit à l'aide de l'estimateur du maximum de vraisemblance $\hat{\theta}^*$.

DÉMONSTRATION. Raisonnons par l'absurde. Il existe une sous-suite $\{n'\}$ de nombres tels que $\Delta\sqrt{n'} \to c\beta/\sqrt{I(\theta)}$, $c < 1$. Comme $\theta^{\pm} = \theta^* \pm \Delta$, il vient

$$1 - \epsilon = \lim_{n'\to\infty} \mathbf{P}_\theta(\theta^- < \theta < \theta^+) = \lim_{n'\to\infty} \mathbf{P}_\theta(|\theta^* - \theta| < \Delta) =$$
$$= \lim_{n'\to\infty} \mathbf{P}_\theta(|\theta^* - \theta|\sqrt{n'} < c\beta/\sqrt{I(\theta)}) \leqslant$$
$$\leqslant \lim_{n\to\infty} \mathbf{P}_\theta(|\hat{\theta}^* - \theta|\sqrt{n} < c\beta/\sqrt{I(\theta)}). \quad (11)$$

La dernière inégalité résulte de ce que l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ est asymptotiquement efficace dans la classe $\tilde{K}^\circ$ des estimateurs asymptotiquement centraux (cf. théorème 25.4). Vu que le dernier membre de (11) est $< 1 - \epsilon$, on obtient une contradiction qui prouve le théorème. ◄

6. Cas vectoriel. La notion d'intervalle de confiance se généralise, dans le cas d'un paramètre vectoriel $\theta \in R^k$ à celle de région de confiance, ou d'ensemble de confiance.

DÉFINITION 4. On dit qu'un sous-ensemble aléatoire *) $\Theta^* = \Theta^*(\epsilon, X)$ d'un espace de paramètres Θ est un *ensemble de confiance au seuil* $1 - \epsilon$ si

$$\mathbf{P}_\theta(\Theta^* \ni \theta) \geqslant 1 - \epsilon. \quad (12)$$

Autrement dit, un ensemble de confiance au seuil $1 - \epsilon$ recouvre la valeur exacte inconnue de θ avec une probabilité $\geqslant 1 - \epsilon$.

*) Dans ce contexte on dira que l'ensemble $\Theta^*(\varepsilon, X)$ est aléatoire si pour chaque t l'ensemble $\{X : t \in \Theta^*(\varepsilon, X)\}$ est mesurable, et par suite, est définie la probabilité (12) (comparer avec le § 3.8).

DÉFINITION 5. On dit qu'un ensemble aléatoire Θ^* est un *ensemble de confiance asymptotique au seuil* $1 - \epsilon$ si $X = [X_\infty]_n \in \mathbf{P}_\theta$ et Θ^* vérifie la relation

$$\liminf_{n\to\infty} \mathbf{P}_\theta(\Theta^* \ni \theta) \geqslant 1 - \epsilon .$$

Les ensembles de confiance « exacts », y compris les ensembles optimaux, feront l'objet du § 8 du chapitre suivant.

Quant aux ensembles de confiance asymptotiques, ils admettent le même principe de construction. Grâce au théorème 4 nous pouvons envisager immédiatement les ensembles de confiance construits à l'aide d'un estimateur du maximum de vraisemblance $\hat{\theta}^*$. On sait que si les conditions (RR) sont remplies et si $X \in \mathbf{P}_\theta$, on a

$$(\hat{\theta}^* - \theta)\sqrt{n}\, I^{1/2}(\theta) \in \Phi_{0,E}.$$

D'où

$$n(\hat{\theta}^* - \theta) I(\theta)(\hat{\theta}^* - \theta)^T \in \mathbf{H}_k,$$
$$n(\hat{\theta}^* - \theta) I(\hat{\theta}^*)(\hat{\theta}^* - \theta)^T \in \mathbf{H}_k.$$

Autrement dit, si h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à k degrés de liberté, alors

$$\lim_{n\to\infty} \mathbf{P}_\theta(n(\theta - \hat{\theta}^*) I(\hat{\theta}^*)(\theta - \hat{\theta}^*)^T < h_\epsilon) = 1 - \epsilon . \qquad (13)$$

L'ensemble de confiance asymptotique Θ^* au seuil $1 - \epsilon$ que nous avons construit est un ellipsoïde de centre $\hat{\theta}^*$ et d'axes définis par la matrice $nI(\hat{\theta}^*)/h_\epsilon$. Pour construire l'ensemble Θ^* il n'est pas obligatoire de calculer la matrice $I(\theta)$. Nous savons que si les conditions (RR) sont remplies et $X \in \mathbf{P}_\theta$, alors

$$L(X, \theta) - L(X, \hat{\theta}^*) \approx -\frac{n}{2}(\theta - \hat{\theta}^*) I(\hat{\theta}^*)(\theta - \hat{\theta}^*)^T.$$

Donc, l'ellipsoïde Θ^* défini dans (13) peut être représenté comme l'ensemble des valeurs θ telles que

$$L(X, \theta) - L(X, \hat{\theta}^*) \geqslant -h_\epsilon/2.$$

Au § 28 on a établi que la limite de la $\mathbf{P}_\theta$-probabilité de cette inégalité (cf. remarque 28.2) est égale à $1 - \epsilon$

Il s'ensuit en particulier que dans le cas scalaire les bornes $\theta^\pm$ de l'intervalle de confiance asymptotique au seuil $1 - \epsilon$ peuvent être définies comme solutions de l'équation

$$L(X, \theta) - L(X, \hat{\theta}^*) = -h_\epsilon/2 = -\beta^2/2.$$

§ 32. Distributions empiriques et intervalles de confiance exacts pour les lois normales

De toutes les distributions énumérées dans le § 2, la distribution normale est la plus fréquente dans les applications. Aussi dans ce paragraphe nous arrêterons-nous spécialement sur la construction d'intervalles de confiance exacts pour les paramètres α et σ^2 de la distribution Φ_{α,σ^2}.

1. Distributions exactes des statistiques $\bar{x}$ et S_0^2. Soient $X \in \Phi_{0,1}$ et $C = |c_{ij}|$ $(i, j = 1, 2, \ldots, n)$ une matrice orthogonale.

Etudions la distribution du vecteur n-dimensionnel $Y = XC$, $Y = (y_1 \ldots \ldots, y_n)$, $y_i = \sum_{j=1}^{n} x_j c_{ji}$.

LEMME 1. *Si C est une matrice orthogonale, alors $Y \in \Phi_{0,1}$, c'est-à-dire que les coordonnées $y_1, \ldots, y_n$ sont des variables aléatoires indépendantes, $y_i \in \Phi_{0,1}$, $i = 1, 2, \ldots, n$.*

DÉMONSTRATION. Soit $t = (t_1, \ldots, t_n)$. Dire que X est normal revient à dire que sa fonction caractéristique est égale à

$$\mathbf{E}e^{itX^T} = e^{-\frac{1}{2}tmt^T},$$

où $m = |m_{ij}|$ est la matrice des moments d'ordre deux, égale ici à une matrice unité E telle que $tEt^T = \sum_{j=1}^{n} t_j^2$,

$$\mathbf{E}e^{itX^T} = e^{-\frac{1}{2}\sum_{j=1}^{n} t_j^2}$$

La fonction caractéristique de la distribution conjointe de $y_1, \ldots, y_n$ (ou de la distribution du vecteur Y) est

$$f(t) = \mathbf{E}e^{itY^T} = \mathbf{E}e^{itC^TX^T}.$$

En faisant le changement $t = uC$ et en remarquant que $CC^T = E$, on obtient

$$f(t) = \mathbf{E}e^{iuCY^T} = \mathbf{E}e^{iuX^T} = e^{-\frac{1}{2}\sum_{j=1}^{n} u_i^2} = e^{-\frac{1}{2}\sum_{j=1}^{n} t_i^2}$$

Ceci exprime que Y admet la même fonction caractéristique, donc la même distribution que X. ◄

Prouvons maintenant une importante proposition pour la suite, appelée *lemme de Fisher.*

LEMME 2. *Supposons toujours que $X \in \Phi_{0,1}$, C est une matrice orthogonale et $Y = (y_1, \ldots, y_n) = XC$. Alors la forme quadratique*

$$T(X) = \sum_{i=1}^{n} x_i^2 - y_1^2 - \ldots - y_r^2$$

ne dépend pas des variables aléatoires $y_1, \ldots, y_r$ et suit une distribution du χ^2 à $n - r$ degrés de liberté : $T(X) \in \mathbf{H}_{n-r}$.

DÉMONSTRATION. Elle coule presque de source puisqu'en appliquant la transformation orthogonale C on obtient

$$T(X) = \sum_{i=1}^{n} y_i^2 - y_1^2 - \ldots - y_r^2 = y_{r+1}^2 + \ldots + y_n^2.$$

Reste seulement à se servir du lemme 1. ◄

Passons maintenant à l'étude de la distribution conjointe des statistiques

$$\bar{x} \text{ et } S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

THÉORÈME 1. *Soit $X \in \Phi_{\alpha,\sigma^2}$. Alors*

1) $(\bar{x} - \alpha)\sqrt{n}/\sigma \in \Phi_{0,1}$,

2) $(n - 1)S_0^2/\sigma \in \mathbf{H}_{n-1}$,

3) *les variables aléatoires $\bar{x}$ et S_0^2 sont indépendantes.*

DÉMONSTRATION. L'assertion 1 est évidente. Il est clair que sans nuire à la généralité on peut admettre que $\alpha = 0$, $\sigma = 1$. On a

$$(n-1)S_0^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

On remarquera que

$$\sqrt{n}\,\bar{x} = \frac{1}{\sqrt{n}} x_1 + \ldots + \frac{1}{\sqrt{n}} x_n$$

et que le vecteur colonne à n dimensions $\begin{pmatrix} 1/\sqrt{n} \\ \ldots \\ 1/\sqrt{n} \end{pmatrix}$ (de norme 1) peut toujours être complété à une matrice orthogonale C. Alors $y_1 = \sqrt{n}\,\bar{x}$ est la première coordonnée de $Y = XC$ et, en appliquant le lemme 2, on trouve que

$$(n-1)S_0^2 = \sum_{i=1}^{n} x_i^2 - y_1^2 \in \mathbf{H}_{n-1}$$

et que les variables aléatoires $(n - 1)S_0^2$ et $y_1 = \sqrt{n}\,\bar{x}$ sont indépendantes. ◄

COROLLAIRE 1. *Soit* $X \in \Phi_{\alpha,\sigma^2}$. *Alors* $t = (\bar{x} - \alpha)\sqrt{n}/S_0 \in T_{n-1}$, *autrement dit, t suit une loi de Student à* $n - 1$ *degrés de liberté.*

Ceci résulte du théorème 1 et de la représentation

$$t = \frac{(\bar{x} - \alpha)\sqrt{n}}{\sigma} \cdot \frac{1}{\sqrt{\frac{1}{n-1} \cdot \frac{(n-1)S_0^2}{\sigma^2}}}. \quad \blacktriangleleft$$

Le théorème 1 d'indépendance de S_0^2 et de $\bar{x}$ peut être renforcé. Il se trouve que $\bar{x}$ ne dépend pas du vecteur $X - \bar{x}$ (c'est-à-dire, ne dépend pas des termes S_0^2). Ceci résulte de la normalité de $\bar{x}$ et $X - \bar{x}$ et de la non-corrélation des variables aléatoires $\bar{x}$ et $x_i - \bar{x}$ qui découle de l'égalité ($\alpha = 0$)

$$E(x_1 - \bar{x})\bar{x} = \frac{1}{n^2}\left[(n - 1)Ex_1^2 - E\left(\sum_{i=2}^{n} x_i\right)^2\right] = 0.$$

2. Construction d'intervalles de confiance exacts pour les paramètres de la distribution normale. Envisageons d'abord deux situations simples.

a) Supposons que $X \in \Phi_{\alpha,\sigma^2}$ et que σ^2 est connue. On se propose de construire un intervalle de confiance correspondant au seuil $1 - \varepsilon$ pour l'estimation de α. La forme de l'intervalle de confiance découle dans ce cas de façon évidente des égalités

$$P(|(\bar{x} - \alpha)\sqrt{n}/\sigma| < \beta) = P(-\sigma\beta/\sqrt{n} < \bar{x} - \alpha < \sigma\beta/\sqrt{n}) = 1 - \varepsilon,$$

où comme précédemment $\beta = \lambda_{\varepsilon/2}$, $\Phi_{0,1}(]-\infty, \lambda_\delta[) = 1 - \delta$, de sorte que

$$\alpha^{\pm}(\varepsilon, X) = \bar{x} \pm \sigma\beta/\sqrt{n}.$$

On propose au lecteur d'appliquer à titre d'exercice la procédure plus formelle développée dans le théorème 31.2 et de se servir de la fonction $G(\alpha, X) = (\bar{x} - \alpha)\sqrt{n}/\sigma \in \Phi_{0,1}$.

b) Supposons α connu. On demande de construire un intervalle de confiance au seuil $1 - \varepsilon$ pour σ^2.

Posons

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2.$$

Il est alors évident que $nS_1^2/\sigma^2 \in H_n$ et par suite

$$P(y_n^- < nS_1^2/\sigma^2 < y_n^+) = H_n(]y_n^-, y_n^+[) = P(nS_1^2/y_n^+ < \sigma^2 < nS_1^2/y_n^-).$$

Les bornes de l'intervalle de confiance cherché seront donc

$$(\sigma^2)^{\pm} = nS_1^2/y_n^{\mp}$$

pour tous $y_n^{\pm}$ tels que $H_n(]y_n^-, y_n^+[) = 1 - \varepsilon$.

Si l'on se sert de la procédure du théorème 31.2, il faut poser $G(\sigma, X) = nS_1^2/\sigma^2 \in \mathbf{H}_n$.

Traitons maintenant le cas où les paramètres α et σ^2 sont tous deux inconnus.

c) Construisons un intervalle de confiance de σ^2 à l'aide de la statistique $G_1(\sigma, X) = (n-1)S_0^2/\sigma^2$. Le théorème 1 affirme que $G_1(\sigma, X) \in \mathbf{H}_{n-1}$. Procédons ensuite comme dans le cas b). Les bornes de l'intervalle de confiance de σ^2 seront

$$(\sigma^2)^{\pm} = (n-1)S_0^2/y_{n-1}^{\mp}.$$

Il est immédiat de voir que dans les cas b) et c) les statistiques $G(\sigma, X)$ et $G_1(\sigma, X)$ sont équidistribuées et par suite conduisent au même intervalle de confiance pour σ^2 si seulement le nombre des observations de c) est supérieur d'une unité à celui de b). De façon plus imagée, dans le cas c) nous « perdons » une observation en raison de l'indétermination supplémentaire introduite par le paramètre inconnu α. Cette observation est en quelque sorte destinée à estimer le paramètre « fantôme » *) α.

d) Construisons maintenant un intervalle de confiance pour α. Servons-nous de la statistique $G_1(\alpha, X) = (\bar{x} - \alpha)\sqrt{n}/S_0$. Le corollaire du théorème 1 nous donne

$$G_1(\alpha, X) \in \mathbf{T}_{n-1}.$$

La fonction $G_1(\alpha, X)$ vérifiant les conditions du théorème 31.2, les raisonnements ultérieurs reprennent *ad litteram* ceux des cas a), b) et c). Les bornes de l'intervalle de confiance sont (pour simplifier on prendra un intervalle symétrique)

$$\alpha^{\pm} = \bar{x} \pm \tau_\epsilon S_0/\sqrt{n},$$

---

*) Il est intéressant de noter que contrairement aux notions intuitives généralement admises il est possible de construire au vu d'une seule observation $x_1 \in \Phi_{\alpha,\sigma^2}$ un intervalle de confiance de σ^2 lorsque α est inconnu. Les raisonnements suivants qui le montrent nous ont été communiqués par L. Bolchev.

Choisissons u tel que $\Phi(1/u) - \Phi(-1/u) = \epsilon$, où $\Phi(x) = \Phi_{0,1}(]-\infty, x[)$. Alors

$$\mathbf{P}(\sigma > u|x_1|) = \mathbf{P}(-\sigma/u < x_1 < \sigma/u) =$$

$$= \mathbf{P}\left(-\frac{1}{u} - \frac{\alpha}{\sigma} < \frac{(x_1 - \alpha)}{\sigma} < \frac{1}{u} - \frac{\alpha}{\sigma}\right) =$$

$$= \Phi\left(\frac{1}{u} - \frac{\alpha}{\sigma}\right) - \Phi\left(-\frac{1}{u} - \frac{\alpha}{\sigma}\right) \leqslant \Phi\left(\frac{1}{u}\right) - \Phi\left(-\frac{1}{u}\right) = \epsilon. \blacktriangleleft$$

où τ_ϵ se déduit de la relation

$$\mathbf{P}(|t_{n-1}| < \tau_\epsilon) = \mathbf{T}_{n-1}(]-\tau_\epsilon, \tau_\epsilon[) = 1 - \epsilon .$$

Remarquons que si S_0 est proche de σ, l'intervalle de confiance sera plus large que celui de a), puisque $\tau_\epsilon > \beta$ (cf. remarque du § 2). Ceci s'explique encore par la présence du paramètre fantôme σ qui est connu dans a).

Les nombres $y^\pm$ qui vérifient la relation

$$\mathbf{P}(G(\theta, X) \in]y^-, y^+[) = 1 - \epsilon ,$$

sont généralement donnés par des tables de statistique mathématique.

Au § 3.8 on montrera que les intervalles de confiance construits dans ce paragraphe sont dans un certain sens les meilleurs.

CHAPITRE 3

THÉORIE DES TESTS D'HYPOTHÈSES

Dans les §§ 1, 2, 3 et 11, on expose la théorie des tests de choix entre un nombre fini (en particulier entre deux) d'hypothèses simples.

Les §§ 4 à 12 sont consacrés à la construction de tests optimaux de choix entre deux hypothèses multiples. On étudie en particulier les tests bayésiens et minimax (§§ 4, 9) et on applique les principes d'exhaustivité, d'absence de biais et d'invariance à la construction des tests uniformément les plus puissants.

Dans les §§ 13 à 17, on développe les méthodes de construction des tests asymptotiquement optimaux.

§ 1. Test de choix entre un nombre fini d'hypothèses simples

1. Position du problème. Notion de test statistique. Test le plus puissant. Dans ce chapitre il sera question de tester des hypothèses concernant une distribution **P** d'un échantillon X. Comme en théorie des estimations il n'y aurait pas de problème si la distribution **P** était connue.

La prise d'une décision concernant l'acceptation ou le rejet d'une hypothèse donnée H doit se baser uniquement sur l'échantillon $X \in \mathbf{P}$ donné et sur une éventuelle information *a priori* sur la distribution **P**.

Donc, pour définir la procédure de prise de décision au vu d'un échantillon X, nous devons définir une application surjective de l'espace des échantillons $\mathscr{X}^n$ sur l'ensemble des hypothèses envisagées. Cette application est généralement appelée *test* ou *critère statistique*. Des définitions exactes seront données plus bas pour diverses situations.

Commençons par le problème le plus simple : le test de choix entre un nombre fini d'hypothèses simples.

DÉFINITION 1. On appellera *hypothèse simple* toute hypothèse définissant de façon unique la distribution de l'échantillon X.

Soient données r distributions $\mathbf{P}_1, \ldots, \mathbf{P}_r$ et supposons que X est issu de l'une d'elles. Le problème est de déterminer cette distribution $\mathbf{P}_j$, $j = 1, 2, \ldots, r$. Chacune des r hypothèses

$$H_j = \{X \in \mathbf{P}_j\} \qquad (1)$$

sera simple et le problème consiste donc à décider entre r hypothèses simples.

De même que dans le chapitre 2, nous traiterons souvent dans ce chapitre le cas paramétrique où X suit une distribution $\mathbf{P}_\theta \in \mathscr{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$. Si la condition ($A_0$) est remplie, les hypothèses simples s'écriront sous la forme $H_j = \{X \in \mathbf{P}_{\theta_j}\}$, où $\theta_1, \ldots, \theta_r$ sont des points fixes de Θ. Le cas (1) peut aussi être traité comme un cas paramétrique avec un ensemble fini $\Theta = \{1, \ldots, r\}$.

Ces considérations montrent qu'il n'y a aucune différence de principe entre l'estimation des paramètres et le test d'hypothèses: dans les deux cas il faut déterminer la valeur inconnue de θ. Il existe tout de même une légère distinction, à savoir que dans le test d'hypothèses les valeurs possibles de θ sont discrètes, et les méthodes liées à la comparaison, disons, des erreurs quadratiques moyennes, qui ont été développées dans le chapitre 2, ne passent pas ici. Nous ferons appel à d'autres critères pour comparer les règles d'acceptation de telle ou telle hypothèse au vu d'un échantillon X.

Le caractère discret de l'ensemble des valeurs possibles de θ fait apparaître un nouvel élément, à savoir qu'il est désormais possible de déterminer exactement, avec une probabilité non nulle, la valeur inconnue θ_i (ou la distribution $\mathbf{P}_{\theta_i}$), alors que dans les problèmes d'estimation de paramètres, la probabilité d'un tel événement était généralement nulle.

DÉFINITION 2. On appelle *test statistique* de choix entre r hypothèses $H_1, \ldots, H_r$ toute application surjective mesurable $\delta : \mathscr{X}^n \to \{H_1, \ldots, H_r\}$.

En d'autres termes, $\delta(X)$ est une « variable » aléatoire prenant les valeurs $H_1, H_2, \ldots, H_r$: si $\delta(X) = H_k$, nous retenons l'hypothèse H_k (c'est-à-dire que nous admettons que $\theta = \theta_k$ dans le cas paramétrique).

L'application $\delta(\cdot)$ s'appelle aussi *règle de décision* ou *fonction de décision*. Il est clair que donner la règle de décision équivaut à définir une partition de l'espace $\mathscr{X}^n$ en r ensembles boréliens disjoints $\Omega_1, \Omega_2, \ldots, \Omega_r$ sur lesquels sont acceptées respectivement les hypothèses $H_1, H_2, \ldots, H_r$.

La qualité d'un test est le plus souvent caractérisée par les probabilités de prise d'une fausse décision:

$$\alpha_i = \alpha_i(\delta) = \mathbf{P}_i(X \notin \Omega_i) = \mathbf{P}_i(\delta(X) \neq H_i).$$

Le nombre α_i représente la probabilité de rejeter l'hypothèse H_i à tort. On l'appelle *probabilité d'erreur* ou *risque de i-ième espèce du test* δ.

Si l'on a réussi à choisir un test δ de telle sorte que tous les nombres α_i soient petits, on admettra en vertu du principe fondamental mentionné dans le § 2.31 qu'il est pratiquement impossible de commettre une erreur en une épreuve et l'on déclarera que l'hypothèse H_k est vraie si $\delta(X) = H_k$. Ce faisant, on se trompera dans $\alpha_i = \mathbf{P}_i(\delta(X) \neq H_i)$ pour cent des cas si l'hypothèse H_i est vraie.

Il est certes souhaitable de tester des hypothèses de façon à rendre minimales les probabilités de ces erreurs. Mais si la taille de l'échantillon X est donnée, il n'est pas possible de rendre les α_i simultanément petits. Tout ce qu'on peut faire, c'est fixer quelques uns d'entre eux et essayer de minimiser les autres.

Nous sommes ainsi amenés à comparer des tests entre eux. Munissons l'ensemble des tests entre les hypothèses $H_1, \dots, H_r$ d'une relation d'ordre partiel.

DÉFINITION 3. On dit qu'un test δ_1 est *meilleur* qu'un test δ_2 si pour tous les $i = 1, 2, \dots, r$

$$\alpha_i(\delta_1) \leqslant \alpha_i(\delta_2)$$

et l'inégalité stricte est réalisée pour un i au moins.

Les tests δ_1 et δ_2 ne sont pas toujours comparables au sens de cette définition. De la même façon il est impossible de comparer deux estimateurs θ_1^* et θ_2^* du point de vue de l'approche de la moyenne quadratique si l'on prend $\mathbf{E}_\theta(\theta^* - \theta)^2$ pour critère de qualité. Pour pouvoir comparer les tests, il faut restreindre l'ensemble des fonctions de décision. Introduisons à cet effet les classes

$$K_{\alpha_1, \dots, \alpha_{r-1}} = \{\delta : \alpha_j(\delta) = \alpha_j \,;\, j = 1, 2, \dots, r-1\}.$$

Les classes $K_{\alpha_1, \dots, \alpha_{r-1}}$ peuvent être munies, elles, d'une relation d'ordre portant sur α_r : plus $\alpha_r(\delta)$ est petit, plus le test est meilleur.

DÉFINITION 4. Un test $\delta_0 \in K_{\alpha_1, \dots, \alpha_{r-1}}$ est dit *le plus puissant dans la classe* $K_{\alpha_1, \dots, \alpha_{r-1}}$ si pour tout δ de cette classe

$$\alpha_r(\delta_0) \leqslant \alpha_r(\delta).$$

On rappelle qu'on a fait une chose semblable au chapitre 2 à propos de la comparaison des estimateurs. Dans ce chapitre on a mis en évidence la classe K_b des estimateurs à biais fixé.

Il existe encore deux approches qui permettent d'ordonner l'ensemble des fonctions de décision à l'aide d'une caractéristique numérique: l'approche bayésienne et l'approche minimax.

Avant de passer à la construction des tests les plus puissants dans les classes $K_{\alpha_1, \dots, \alpha_{r-1}}$, considérons ces deux approches.

2. Approche bayésienne. Cette approche admet que la distribution $\mathbf{P}_j$ de l'échantillon X a été choisie de façon aléatoire. Dans ce cas les hypothèses $H_j = \{X \in \mathbf{P}_j\}, j = 1, \dots, r$, sont des événements aléatoires dont nous désignerons les probabilités par

$$\mathbf{Q}(H_j) = q(j),$$

de sorte que $\mathbf{Q}$ est une distribution *a priori* sur l'ensemble des hypothèses $\{H_1, ..., H_r\}$ et $q(j)$, $j = 1, ..., r$, sont les probabilités *a priori* de ces hypothèses (comparer avec le § 2.11). Il devient plus simple dans ces conditions de comparer des tests, puisque nous pouvons définir la *probabilité moyenne d'erreur*, ou *risque moyen*, $\alpha_{\mathbf{Q}}(\delta)$ du test δ :

$$\alpha_{\mathbf{Q}}(\delta) = \sum_{j=1}^{r} \mathbf{Q}(H_j)\mathbf{P}_j(\delta(X) \neq H_j) = \sum_{j=1}^{r} q(j)\alpha_j(\delta) \tag{2}$$

et par conséquent, ordonner totalement l'ensemble des tests par rapport à $\alpha_{\mathbf{Q}}(\delta)$.

DÉFINITION 5. Le test $\delta = \delta_{\mathbf{Q}}$ qui minimise $\alpha_{\mathbf{Q}}(\delta)$ s'appelle *test bayésien associé à la distribution a priori* $\mathbf{Q}$.

Soit remplie la condition (A_μ), c'est-à-dire que les distributions $\mathbf{P}_j$ admettent des densités $f_j(x)$ par rapport à une mesure σ-finie μ. La fonction $f_j(X) = \prod_{i=1}^{n} f_j(x_i)$ sera appelée comme précédemment *fonction de vraisemblance*.

La fonction $f(x) = \sum q(j)f_j(x)$ est la densité de la distribution de X par rapport à la mesure μ^n, et $q(j)f_j(x)$ la densité de la distribution conjointe du couple (θ, X) dans lequel le numéro θ de l'hypothèse est choisi de façon aléatoire.

Si donc est donné l'échantillon X, on peut, dans le cas bayésien, construire une *distribution a posteriori* $\mathbf{Q}_X$ des hypothèses H_j (la mesure λ du § 2.11 est ici une mesure cardinale) à l'aide de la formule de Bayes

$$\mathbf{Q}_X(H_k) = q(k \mid X) = \frac{q(k)f_k(X)}{f(X)}. \tag{3}$$

Ceci est la *distribution conditionnelle de* θ *par rapport à* X.

Par $\mathbf{E}$ on désignera l'espérance mathématique associée à la distribution $\mathbf{P}$ du couple (θ, X).

THÉORÈME 1. 1) *Le risque* $\alpha_{\mathbf{Q}}(\delta)$ *de tout test* δ *vérifie l'inégalité*

$$\alpha_{\mathbf{Q}}(\delta) \geqslant 1 - \mathbf{E} \max_j q(j \mid X). \tag{4}$$

2) *Pour qu'un test* $\delta = \delta_{\mathbf{Q}}$ *soit bayésien pour une distribution a priori* $\mathbf{Q}$, *il est nécessaire et suffisant que pour* $\mathbf{P}$*-presque toutes les valeurs de* X *il vérifie les relations*

$$\delta(X) = H_k \quad si \quad q(k \mid X) = \max_j q(j \mid X). \tag{5}$$

L'égalité est réalisée dans (4) *pour* $\delta = \delta_{\mathbf{Q}}$.

Signalons que le second membre de (4) est indépendant de δ.

DÉMONSTRATION. Soit donné un test δ. Considérons l'événement D_δ qui consiste en ce que le test δ conduise à prendre une fausse décision :

$$D_\delta = \bigcup_{j=1}^{r} \{\theta = j, \delta(X) \neq H_j\}.$$

Il est alors évident que $\alpha_Q(\delta) = \mathbf{P}(D_\delta)$ et la notation (2) est le résultat d'une moyennisation : d'abord par rapport à X pour $\theta = j$ et ensuite par rapport à θ. Nous pouvons écrire $\alpha_Q(\delta)$ sous une autre forme: prendre la moyenne par rapport à θ pour X fixe et ensuite par rapport à X :

$$\alpha_Q(\delta) = \int \mathbf{P}(D_\delta | X = x) f(x) \mu(dx) =$$

$$= \mathbf{EP}(D_\delta | X) = \mathbf{E} \sum_{j=1}^{r} \mathbf{P}(\theta = j, \delta(X) \neq H_j | X).$$

Puisque $\delta(X)$ est mesurable par rapport à X, on a

$$\mathbf{P}(\theta = j, \delta(X) \neq H_j | X) = I_{\{\delta(X) \neq H_j\}} \mathbf{P}(\theta = j | X) =$$

$$= (1 - I_{\{\delta(X) = H_j\}}) q(j | X).$$

D'où

$$\alpha_Q(\delta) = 1 - \mathbf{E} \sum_{j=1}^{r} q(j | X) I_{\{\delta(X) = H_j\}} \geqslant 1 - \mathbf{E} \max_j q(j | X).$$

Ce qui prouve la première partie du théorème.

La suffisance de la deuxième proposition du théorème découle de toute évidence de la première proposition, puisque la borne inférieure établie pour $\alpha_Q(\delta)$ est atteinte pour le test δ_Q défini dans (5). Il est évident que $\alpha_Q(\delta_Q)$ ne change pas lorsque $\delta_Q(X)$ varie sur un ensemble de $\mathbf{P}$-probabilité nulle.

La nécessité de la deuxième proposition se prouve de façon aussi simple. En effet, supposons que $\delta = \delta_Q$ est un test bayésien et que $\delta(X) = H_k$, $q(k | X) < q(l | X) = \max_j q(j | X)$ pour $X \in A$, $\mathbf{P}(A) > 0$. Alors pour le test $\delta_1(X)$ qui ne diffère de $\delta(X)$ que sur l'ensemble A : $\delta_1(X) = H_l$ pour $X \in A$, on obtient

$$\mathbf{P}(D_{\delta_1}; A) = \mathbf{P}(A) - \mathbf{E}\left[\sum_j q(j | X) I_{\{\delta_1(X) = H_j\}}; A\right] =$$

$$= \mathbf{P}(A) - \mathbf{E}[q(l | X); A] < \mathbf{P}(A) - \mathbf{E}[q(k | X); A] = \mathbf{P}(D_\delta; A);$$

$$\mathbf{P}(D_{\delta_1}) < \mathbf{P}(D_\delta) = \mathbf{P}(D_{\delta_Q}).$$

Nous avons obtenu une contradiction. ◀

Signalons maintenant que la notation (5) ne définit pas entièrement le test δ_Q : elle n'indique pas clairement quelle hypothèse il faut retenir si deux valeurs ou plus de $q(j \mid X)$ sont maximales. Il s'agit visiblement de définir la fonction $\delta_Q(X)$ sur les frontières

$$\Gamma_k = \{x \in \mathscr{X}^n : q(k)f_k(x) = \max_{j \neq k} q(j)f_j(x)\}$$

des ensembles

$$\tilde{\Omega}_k^Q = \{x \in \mathscr{X}^n : q(k)f_k(x) > \max_{j \neq k} q(j)f_j(x)\} \qquad (6)$$

dans lesquels, en vertu de (5), il faut accepter l'hypothèse H_k d'après le test δ_Q.

Donc $\tilde{\Omega}_k^Q$ est l'« intérieur » de la région

$$\Omega_k^Q = \{x \in \mathscr{X}^n : \delta_Q(x) = H_k\}$$

d'acceptation de l'hypothèse H_k, et il nous faut, en plus de (6), déterminer les seuls points de Γ_k qui appartiennent à Ω_k^Q. Or des raisonnements précédents il ressort que ce problème peut être résolu de façon assez élémentaire: nous pouvons adjoindre les points de Γ_k à n'importe lequel des domaines « adjacents » $\tilde{\Omega}_j^Q$; ce faisant nous obtiendrons la même valeur de $\alpha_Q(\delta)$, puisque (5) aura lieu. Plus exactement, si $A \subset \Gamma_{k_1} \cap \dots \cap \Gamma_{k_l}$ et que $X \in A$, peu importe, en vertu du test bayésien, laquelle des hypothèses $H_{k_1}, \dots, H_{k_l}$ est acceptée. Nous pouvons même prendre une décision de façon aléatoire : choisir l'hypothèse H_{k_i}, $i = 1, \dots, l$, avec la probabilité p_{k_i}, $\sum_{i=1}^{l} p_{k_i} = 1$. La valeur de $\alpha_Q(\delta)$ ne change pas.

Nous sommes conduits ici à une notion plus générale de test randomisé, qui est très utile.

DÉFINITION 6. On appelle *test randomisé* entre les hypothèses $H_1, \dots$ $\dots, H_r$ toute application surjective mesurable $\pi : \mathscr{X}^n \to R^{(r)}$, où $R^{(r)}$ est l'ensemble des vecteurs $(\pi_1, \dots, \pi_r)$, $\pi_i \geqslant 0$, $\sum_{i=1}^{r} \pi_i = 1$.

Un test randomisé associe à tout $x \in \mathscr{X}^n$ une distribution de probabilités $\pi(x) = (\pi_1(x), \dots, \pi_r(x))$ sur l'ensemble $\{H_1, \dots, H_r\}$ et la décision finale concernant l'acceptation de l'hypothèse est prise au hasard (indépendamment de X une fois que les probabilités $\pi_i(X)$ ont été définies).

Un test ordinaire est visiblement un cas particulier d'un test randomisé lorsque toutes les probabilités π_i sont nulles à l'exception d'une seule qui est égale à 1. De tels tests seront dits *non randomisés* ou *déterministes*.

Le risque de i-ième espèce $\alpha_i(\pi)$ d'un test randomisé se définit de façon

analogue :

$$\alpha_i(\pi) = \mathbf{P}_i(\textit{de rejeter } H_i) = 1 - \mathsf{E}_i\pi_i(X).$$

Du point de vue bayésien la minimisation de

$$\alpha_{\mathbf{Q}}(\pi) = \sum_{j=1}^{r} q(j)\alpha_j(\pi)$$

se traite de façon tout à fait analogue. Si par θ on désigne comme précédemment le numéro d'une hypothèse choisie de façon aléatoire, de distribution *a priori* $\mathbf{Q}$, en sorte que $\mathbf{Q}(\theta = j) = q(j)$, alors

$$\alpha_{\mathbf{Q}}(\pi) = 1 - \sum_{j=1}^{r} q(j)\mathsf{E}_j\pi_j(X) = 1 - \mathsf{E}\pi_\theta(X) = 1 - \mathsf{E}\mathsf{E}(\pi_\theta(X)|X) =$$

$$= 1 - \mathsf{E}\sum_{j=1}^{r} q(j|X)\pi_j(X) \geqslant 1 - \mathsf{E}\max_j q(j|X).$$

Nous avons ainsi obtenu la même borne inférieure pour $\alpha_{\mathbf{Q}}(\pi)$ que dans le cas d'un test non randomisé. Ceci exprime qu'un élargissement de la classe des tests n'améliore pas ici la valeur de $\alpha_{\mathbf{Q}}(\delta)$. Bien plus, la plus petite valeur est atteinte sur un test non randomisé $\delta_{\mathbf{Q}}$. Mais le nombre des tests *randomisés* bayésiens $\pi^{\mathbf{Q}}$, c'est-à-dire des tests tels que $\alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}) = \alpha_{\mathbf{Q}}(\delta_{\mathbf{Q}})$, sera bien plus élevé que celui des tests non randomisés, puisque sur l'ensemble

$$\Gamma_{k_1, \ldots, k_l} = \bigcap_{i=1}^{l} \Gamma_{k_i} \bigcap_{j \neq k_1, \ldots, k_l} \bar{\Gamma}_j,$$

où $\bar{\Gamma} = \mathscr{X}^n \setminus \Gamma$, nous pouvons prendre pour $\pi^{\mathbf{Q}}(x)$ n'importe quel vecteur du sous-ensemble $R_{k_1, \ldots, k_l} \subset R^{(r)}$ des vecteurs π dont sont non nulles les seules coordonnées d'indices $k_1, \ldots, k_l$. Il est évident que R_k est composé d'un seul vecteur e_k dont la k-ième coordonnée est égale à 1 et les autres à 0, et l'on doit poser

$$\pi^{\mathbf{Q}}(x) = e_k \quad \text{pour} \quad x \in \tilde{\Omega}_k^{\mathbf{Q}}.$$

Puisque les relations ci-dessus sont, aux valeurs près de $\pi^{\mathbf{Q}}(x)$ sur un ensemble $\mathbf{P}$-négligeable, nécessaires et suffisantes pour que

$$\alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}) = \alpha_{\mathbf{Q}}(\delta_{\mathbf{Q}}) = 1 - \mathsf{E}\max_j q(j|X),$$

nous pouvons en plus du théorème 1 formuler la proposition suivante.

THÉORÈME 1A. 1) *Pour tout test randomisé,*

$$\alpha_{\mathbf{Q}}(\pi) \geqslant 1 - \mathbf{E} \max_j q(j \mid X).$$

2) *Pour qu'un test $\pi^{\mathbf{Q}}$ soit bayésien, il est nécessaire et suffisant que*

$$\pi^{\mathbf{Q}}(x) = e_k \quad si \quad x \in \tilde{\Omega}_k^{\mathbf{Q}}, \tag{7}$$

$$\pi^{\mathbf{Q}}(x) \in R_{k_1, \ldots, k_l} \quad si \quad x \in \Gamma_{k_1, \ldots, k_l}$$

pour **P**-*presque toutes les valeurs de* x.

3) *Pour tout* $g_j \geqslant 0, j = 1, \ldots, r, \sum_{j=1}^{r} g_j = 1$, *on a l'inégalité*

$$\alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}) = \sum_{j=1}^{r} q(j) \alpha_j(\pi^{\mathbf{Q}}) \leqslant \sum_{j=1}^{r} q(j)(1 - g_j). \tag{8}$$

Si $\min_j q_j > 0$ *et les* $f_i(x)$ *ne sont pas toutes confondues, c'est-à-dire qu'il existe des valeurs* k *et* j *et un ensemble* A, $\mathbf{P}(A) > 0$, *sur lequel* $f_k(x) \neq$ $\neq f_j(x)$, *l'inégalité* (8) *est stricte.*

REMARQUE 1. De (8) il s'ensuit que

$$\alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}) \leqslant 1 - \max_j q(j). \tag{9}$$

Le second membre est l'expression du risque d'un test qui conduit à choisir H_k si $q(k) = \max_j q(j)$ (ce test est bayésien dans la classe des tests indépendants de X).

DÉMONSTRATION du théorème 1A. Nous avons déjà prouvé les deux premières propositions. Pour établir la dernière, il suffit de comparer le test bayésien $\pi^{\mathbf{Q}}$ au test $\pi^{\circ}(X) = g = (g_1, \ldots, g_r)$ qui est indépendant de X et pour lequel de toute évidence $\alpha_j(\pi^{\circ}) = 1 - g_j$,

$$\alpha_{\mathbf{Q}}(\pi^{\circ}) = \sum_{j=1}^{r} q(j)(1 - g_j) \geqslant \alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}).$$

Si l'égalité est réalisée dans (8), le test $\pi^{\circ}(X) = g = \text{const}$ sera bayésien. D'après la deuxième proposition du théorème, ceci n'est possible que lorsque $q(1 \mid X) = \ldots = q(r \mid X)$ **P**-presque partout. Ce qui à son tour n'est possible que si $f_1(X) = \ldots = f_r(X)$ **P**-presque partout, $q(1) = \ldots$ $\ldots = q(r)$. ◄

Ainsi, l'introduction des tests randomisés ne permet pas de diminuer le risque α_Q, mais d'accroître le nombre des tests et notamment le nombre des tests bayésiens π^Q. Cette circonstance est parfois utile.

Dans la suite, par test on entendra un test randomisé π.

3. Approche minimax. Si dans l'approche bayésienne nous avons apprécié la qualité d'un test à l'aide de la moyenne $\alpha_Q(\pi) = \sum q(j)\alpha_j(\pi)$, dorénavant nous allons comparer les valeurs maximales

$$\alpha(\pi) = \max_j \alpha_j(\pi) = \max_Q \alpha_Q(\pi).$$

Il est évident que ceci permet aussi d'ordonner l'ensemble des tests.

DÉFINITION 7. On dit qu'un test $\pi = \bar{\pi}$ est *minimax* si

$$\alpha(\bar{\pi}) = \min_\pi \alpha(\pi).$$

La proposition suivante est identique au théorème 2.11.2.

THÉORÈME 2. *Supposons qu'il existe un test bayésien* $\bar{\pi}$ *(associé à une distribution a priori* $\bar{Q}$*) tel que*

$$\alpha_1(\bar{\pi}) = \ldots = \alpha_r(\bar{\pi}). \tag{10}$$

Alors $\bar{\pi}$ *est un test minimax.*

DÉMONSTRATION. Désignons par $\bar{q}(j)$ les probabilités *a priori* associées à $\bar{Q}$. Pour tout test π, on a alors

$$\alpha(\pi) \geqslant \sum_{j=1}^{r} \bar{q}(j)\alpha_j(\pi) \geqslant \sum_{j=1}^{r} \bar{q}(j)\alpha_j(\bar{\pi}) = \max_j \alpha_j(\bar{\pi}) = \alpha(\bar{\pi}). \blacktriangleleft$$

La distribution $Q = \{q(j)\}$ associée au test π est dite *la plus défavorable* (cf. § 2.11). Ceci est lié au fait que pour $Q = Q$

$$\max_Q \alpha_Q(\pi^Q) = \max_Q \min_\pi \alpha_Q(\pi),$$

de sorte que le test minimax (10) est le test bayésien de plus grand risque. La démonstration de ce fait figure dans le chapitre V. On y montre notamment qu'il existe toujours une distribution la plus défavorable et un test minimax.

Signalons toutefois que contrairement au cas bayésien les tests *non randomisés* minimax n'existent pas toujours. En effet, les frontières Γ_k des ensembles $\tilde{\Omega}_k^Q$ (cf. (6)) sont de probabilité non nulle $\mathbf{P}_k(X \in \Gamma_k) > 0$ et par suite, les $\alpha_k(\delta_Q)$ peuvent varier par sauts lorsque Q varie continûment. Ceci exprime à son tour que $r-1$ équations $\alpha_1(\delta_Q) = \ldots = \alpha_r(\delta_Q)$ en $r-1$

inconnues $q(1), \ldots, q(r-1)$ $\left(q(r) = 1 - \sum_{j=1}^{r-1} q(j)\right)$ peuvent ne pas admettre de solution. Mais dans la classe des tests bayésiens randomisés, il existe toujours un test minimax. A titre d'illustration nous étudierons en détail cette question pour $r = 2$ dans le paragraphe suivant.

Nous avons ainsi trouvé la forme explicite des tests bayésiens et avons établi qu'ils pouvaient être utilisés à la construction de tests minimax. Il s'avère qu'il est possible de construire de façon analogue les tests les plus puissants dans les classes $K_{\alpha_1, \ldots, \alpha_{r-1}}$ introduites dans le n° 1.

4. Tests les plus puissants. La définition d'un test le plus puissant non randomisé a été donnée au n° 1. Il est commode de généraliser cette définition à la classe des tests randomisés. Supposons que $K_{\alpha_1, \ldots, \alpha_{r-1}}$ désigne, comme dans le n° 1, la classe des tests *randomisés* de risques de j-ième espèce donnés, $j = 1, \ldots, r-1$:

$$K_{\alpha_1, \ldots, \alpha_{r-1}} = \{\pi : \alpha_j(\pi) = \alpha_j \,; j = 1, \ldots, r-1\}.$$

DÉFINITION 8. Un test $\pi_0 \in K_{\alpha_1, \ldots, \alpha_{r-1}}$ est dit *le plus puissant dans* $K_{\alpha_1, \ldots, \alpha_{r-1}}$ si pour tout $\pi \in K_{\alpha_1, \ldots, \alpha_{r-1}}$ on a

$$\alpha_r(\pi_0) \leqslant \alpha_r(\pi).$$

THÉORÈME 3. *Supposons qu'il existe une distribution* $\mathbf{Q} = \{q(1), \ldots$ $\ldots, q(r)\}$ *telle que*

$$\alpha_j(\pi^{\mathbf{Q}}) \equiv 1 - \mathsf{E}_j \pi_j^{\mathbf{Q}}(X) = \alpha_j, \quad j = 1, \ldots, r-1 \tag{11}$$

(en fait nous avons $r-1$ équations en $r-1$ inconnues $q(1), \ldots, q(r-1)$). *Le test bayésien* $\pi^{\mathbf{Q}}$ *défini dans* (6), (7) *est alors le plus puissant dans la classe* $K_{\alpha_1, \ldots, \alpha_{r-1}}$.

DÉMONSTRATION. Par définition d'un test bayésien

$$\alpha_{\mathbf{Q}}(\pi^{\mathbf{Q}}) \leqslant \alpha_{\mathbf{Q}}(\pi).$$

Ceci exprime que pour $\pi \in K_{\alpha_1, \ldots, \alpha_{r-1}}$, on aura

$$\sum_{j=1}^{r} q(j)\alpha_j(\pi^{\mathbf{Q}}) \leqslant \sum_{j=1}^{r-1} q(j)\alpha_j + q(r)\alpha_r(\pi).$$

Or $\alpha_j(\pi^{\mathbf{Q}}) = \alpha_j$ pour $j \leqslant r-1$, donc $\alpha_r(\pi^{\mathbf{Q}}) \leqslant \alpha_r(\pi)$. ◄

Les équations (11) n'admettent pas toujours une solution dans la classe des tests non randomisés δ pour la même raison que dans la recherche des tests minimax. La situation est fondamentalement différente dans la classe des tests randomisés. Cette circonstance sera illustrée dans le paragraphe suivant.

Exhibons maintenant un exemple assez répandu de test entre un nombre fini d'hypothèses simples.

EXEMPLE 1. Supposons que l'hypothèse H_1 exprime qu'un patient venu consulter son médecin est sain et H_k, qu'il souffre d'une maladie A_k, $k \geqslant 2$. La tâche du médecin est de choisir une hypothèse H_j au vu des observations (que l'on peut représenter sous forme d'un vecteur $\mathbf{x}_1 = (\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1s})$ qui est un échantillon multidimensionnel X de taille un). Les maladies A_k sont fixées pour que les hypothèses H_k soient simples et de ce fait définissent complètement la distribution de X. Si le médecin accepte l'hypothèse H_k, $k \geqslant 2$, alors que H_1 est vraie, il commet une erreur de première espèce. Si au contraire il reconnaît qu'un malade (H_k) est sain (H_1) il commet une erreur de deuxième espèce. Il est clair que les « effets » de ces erreurs peuvent être fondamentalement différents.

De ce qui précède il s'ensuit que pour construire la meilleure règle de décision, il faut connaître la distribution du vecteur des observations $(\mathbf{x}_{11}, \ldots, \mathbf{x}_{1s})$ pour les personnes saines et pour les personnes atteintes des maladies A_k (il faut disposer à cet effet d'importantes données statistiques des examens médicaux). Il va de soi que l'essentiel du problème réside dans le choix de s et des observations $(\mathbf{x}_{11}, \mathbf{x}_{12}, \ldots, \mathbf{x}_{1s})$, ce qui dépend dans une grande mesure de l'art et de l'expérience du médecin.

Si le vecteur $(\mathbf{x}_{11}, \ldots, \mathbf{x}_{1s})$ est choisi assez correctement, les théorèmes 1, 2 et 3 nous indiquent une voie directe d'algorithmisation du diagnostic des maladies.

§ 2. Test de choix entre deux hypothèses simples

Dans ce paragraphe on s'arrêtera plus en détail sur le cas particulier où l'on éprouve r $(=2)$ hypothèses simples.

Ces hypothèses jouent souvent un rôle non symétrique comme, disons, dans l'exemple 1.1. C'est pourquoi l'une de ces hypothèses, par exemple H_1, est appelée *hypothèse de base*, les autres, *hypothèses concurrentes* ou *alternatives* ou encore *contre-hypothèses*. Le risque de première espèce $\alpha_1(\delta)$ d'un test δ s'appelle aussi dans ce cas *dimension* du test et le nombre $1 - \alpha_1(\delta)$ *niveau* du test. Le nombre $\beta(\delta) = 1 - \alpha_2(\delta)$ est dit *puissance* du test δ.

La région $\Omega_2 \subset \mathscr{X}^n$ d'acceptation de l'hypothèse H_2 d'un test non randomisé δ dans le cas où $r = 2$ s'appelle *région critique*. La probabilité $\mathbf{P}_2(X \in \Omega_2)$ d'accès à cette région, lorsque H_2 est vraie, sera égale à la puissance $\beta(\delta)$ du test. D'où l'origine de la dénomination de « test le plus puissant » pour le test δ sur lequel $\beta(\delta)$ atteint son maximum pour un niveau donné.

Signalons maintenant que pour $r = 2$ tout test, qu'il soit randomisé ou

non, peut être caractérisé par une fonction numérique. En effet, un test randomisé $\pi(x)$ est entièrement défini par les valeurs de ses r coordonnées $(\pi_1(x), \ldots, \pi_r(x))$. Mais comme $\sum \pi_j(x) = 1$, il suffit dans le cas $r = 2$ de se donner une seule fonction, disons $\pi_2(x)$. Cette fonction définit la probabilité d'accepter l'hypothèse alternative H_2. Nous la désignerons par $\pi(x)$ et l'appellerons *fonction critique* du test π. Il est évident que pour les tests non randomisés, la fonction $\pi(x)$ ne prend que les valeurs 0 et 1 ; dans le cas général, $0 \leqslant \pi(x) \leqslant 1$.

Le risque $\alpha_1(\pi)$ du test π (ou δ) et sa puissance $\beta(\pi)$ s'expriment en fonction de $\pi(x)$ de la manière suivante :

$$\alpha_1(\pi) = \mathsf{E}_1\pi(X), \quad \beta(\pi) = 1 - \alpha_2(\pi) = \mathsf{E}_2\pi(X).$$

Désignons par Z le rapport de vraisemblance

$$Z = Z(x) = f_2(x)/f_1(x)$$

que nous étudierons pour les seules valeurs de x pour lesquelles il est défini, c'est-à-dire pour les x tels que $f_1(x) + f_2(x) > 0$.

THÉORÈME 1. 1) *Soit* $c = q(1)/q(2)$, *où* $\mathbf{Q} = (q(1), q(2))$, $q(2) = 1 - q(1)$, *est une distribution a priori donnée. Alors le test* $\pi_{c,p}$ *de fonction critique*

$$\pi_{c,p}(x) = \begin{cases} 1 & si \quad Z(x) > c, \\ p(x) & si \quad Z(x) = c, \\ 0 & si \quad Z(x) < c, \end{cases} \tag{1}$$

est bayésien pour la distribution $\mathbf{Q}$ $(\pi_{c,p} = \pi^{\mathbf{Q}})$ *quelle que soit la fonction mesurable* $p(x)$, $0 \leqslant p(x) \leqslant 1$.

Les paramètres $\alpha_1(\pi_{c,p})$ *et* $\alpha_2(\pi_{c,p})$ *du test* $\pi_{c,p}$ *vérifient l'inégalité*

$$\sum_{j=1}^{2} q(j)\alpha_j(\pi_{c,p}) \leqslant \sum_{j=1}^{2} q(j)(1 - g_j) \tag{2}$$

pour tous $g_j \geqslant 0$, $g_1 + g_2 = 1$.

2) *Pour tout* $\epsilon > 0$ *tel que* $\mathbf{P}_1(Z > 0) \geqslant \epsilon$ *il existe un nombre* $c > 0$ *et une fonction* $p(x) \equiv p = \text{const}$ *tels que* $\pi_{c,p} \in K_\epsilon = \{\pi : \alpha_1(\pi) = \epsilon\}$ *et* $\pi_{c,p}$ *est le plus puissant dans* K_ϵ. *Les nombres c et p sont solutions de l'équation*

$$\alpha_1(\pi_{c,p}) \equiv \mathsf{E}_1\pi_{c,p}(X) \equiv \mathbf{P}_1(Z(X) > c) + p\mathbf{P}_1(Z(X) = c) = \varepsilon. \tag{3}$$

De plus, la puissance $\beta(\pi_{c,p}) = 1 - \alpha_2(\pi_{c,p})$ *du test* $\pi_{c,p}$ *vérifie l'inégalité*

$$\beta(\pi_{c,p}) \geqslant \epsilon. \tag{4}$$

Si la relation $f_2(x) = f_1(x)$ $[\mu]$*-presque partout n'est pas réalisée, les inégalités* (4) *et* (2) *sont strictes pour* $0 < q_1 < 1$.

Le test $\pi_{c,p}$ *minimise le risque de première espèce* $\alpha_1(\pi)$ *dans la classe* K *des tests* π *de risque de deuxième espèce donné* : $K = \{\pi : \alpha_2(\pi) = \alpha_2(\pi_{c,p})\}$.

3) *Il existe un nombre* $c > 0$ *et une fonction* $p(x) \equiv p =$ const *tels que le test* $\pi_{c,p}$ *est minimax. Les nombres* c *et* p *se déduisent de l'équation* $\alpha_1(\pi_{c,p}) = \alpha_2(\pi_{c,p})$ *ou, ce qui est équivalent, de l'équation*

$$\mathbf{P}_1(Z(X) > c) + \mathbf{P}_2(Z(X) > c) + p[\mathbf{P}_1(Z(X) = c) + \mathbf{P}_2(Z(X) = c)] = 1. \tag{5}$$

Il est évident que si la $\mathbf{P}_1$-distribution de $Z(X)$ est continue, c'est-à-dire que $\mathbf{P}_1(Z(X) = c) = 0$ pour tous les $c \geqslant 0$, alors on peut poser $p \equiv 1$ ou $p \equiv 0$ dans les deux dernières propositions du théorème.

Remarquons encore que

$$\mathbf{P}_1(Z(X) = c) = \int_{Z(x)=c} f_1(x)\mu^n(dx) = \int_{Z(x)=c} \frac{f_2(x)}{c}\,\mu^n(dx) = \frac{1}{c}\,\mathbf{P}_2(Z(X) = c),$$

de sorte que la continuité sur $]0, \infty[$ de la $\mathbf{P}_1$-distribution de Z entraîne celle de la $\mathbf{P}_2$-distribution de Z.

Le test $\pi_{c,p}$ basé sur le rapport de vraisemblance Z s'appelle *test du rapport de vraisemblance*.

Le théorème 1 montre que *tous les tests optimaux sont des tests du rapport de vraisemblance*.

La deuxième proposition du théorème 1 s'appelle *lemme de Neyman-Pearson*. Si la condition $\mathbf{P}_1(Z > 0) \geqslant \epsilon$ n'est pas remplie dans ce théorème, c'est-à-dire si $\mathbf{P}_1(Z = 0) = 1 - \delta$, $\delta < \epsilon$, le test le plus puissant $\pi(x) = I_{\{Z(x)>0\}}$ aura alors une puissance égale à 1 et un risque $\delta < \epsilon$. Si les supports des distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ sont disjoints, alors $Z = 0$ sur l'ensemble des x tels que $f_1(x) > 0$ et par suite, $\mathbf{P}_1(Z > 0) = 0$. Dans ce cas, les hypothèses H_1 et H_2 sont discernables au vu d'une seule observation avec des probabilités d'erreur nulles, i.e. sont discernables de façon déterministe.

DÉMONSTRATION du théorème 1. La première proposition est une conséquence directe du théorème 1.1A.

Utilisons le théorème 1.3 pour prouver la deuxième. Montrons tout d'abord que l'équation (3) admet toujours une solution en c et p. Il est évident que la fonction $\varphi(c) = \mathbf{P}_1(Z > c)$ est décroissante sur $[0, \infty[$. La variable aléatoire Z est propre par rapport à la distribution $\mathbf{P}_1$, c'est-à-dire

que

$$\varphi(c) = \mathbf{P}_1(Z > c) =$$
$$= \int\limits_{Z(x)>c} f_1(x)\mu^n(dx) < \frac{1}{c} \int\limits_{Z(x)>c} f_2(x)\mu^n(dx) = \frac{1}{c}\, \mathbf{P}_2(Z > c) \to 0$$

lorsque $c \to \infty$. Vu que $\varphi(0) \geqslant \epsilon$ par hypothèse, il existe un $c_\epsilon \in \,]0, \infty[$ tel que

$$\varphi(c_\epsilon - 0) \geqslant \epsilon, \quad \varphi(c_\epsilon) \leqslant \epsilon. \tag{6}$$

Si dans (3) on pose $c = c_\epsilon$ et $\Delta_\epsilon = \varphi(c_\epsilon - 0) - \varphi(c_\epsilon)$, on obtient

$$\alpha_1(\pi_{c_\epsilon, p}) = \varphi(c_\epsilon) + p\Delta_\epsilon.$$

Il est évident qu'en vertu de (6) on peut toujours choisir un $p \in [0, 1]$ tel que *) $\varphi(c_\epsilon) + p\Delta_\epsilon = \epsilon$.

Nous pouvons désormais procéder comme dans la démonstration du théorème 1.3. Posons $q(1) = q_\epsilon = c_\epsilon/(c_\epsilon + 1)$ et figeons le p choisi. Alors le test $\pi_{c_\epsilon, p}$ sera un test bayésien associé à la distribution $\mathbf{Q}_\epsilon = (q_\epsilon, 1 - q_\epsilon)$, et dans le même temps $\alpha_1(\pi_{c_\epsilon, p}) = \epsilon$. Ceci exprime en vertu du théorème 1.3 que $\pi_{c_\epsilon, p}$ est le plus puissant dans K_ϵ.

Si $\pi(x) \equiv \epsilon$, on obtient

$$\pi \in K_\epsilon, \quad \alpha_2(\pi_{c_\epsilon, p}) \leqslant \alpha_2(\pi) \equiv 1 - \epsilon, \quad \beta(\pi_{c_\epsilon, p}) \geqslant \epsilon.$$

Ceci n'est autre que l'inégalité (2) ((1.8)) pour $g_2 = \epsilon$. Donc, si la relation $f_2(x) = f_1(x)$ $[\mu]$-presque partout n'est pas remplie, ces inégalités seront strictes. La proposition du théorème qui dit que $\alpha_1(\pi)$ est minimisé par le test $\pi_{c, p}$ dans la classe $K = \{\pi : \alpha_2(\pi) = \alpha_2(\pi_{c, p})\}$ résulte des raisonnements ci-dessus et de la symétrie par rapport aux hypothèses H_1 et H_2 de la position du problème dans la première proposition du théorème.

Pour démontrer la troisième proposition du théorème 1, il faut se servir du théorème 1.2. A cet effet il nous faut vérifier seulement que l'équation $\alpha_1(\pi_{c, p}) = \alpha_2(\pi_{c, p})$ admet une solution en c et p. Cette équation peut être mise sous la forme

$$\mathsf{E}_1\pi_{c, p}(X) = 1 - \mathsf{E}_2\pi_{c, p}(X)$$

ou, ce qui est équivalent, sous la forme (5). La solubilité de cette équation s'établit exactement comme celle de l'équation (3). On remarquera seulement que toujours $\mathbf{P}_1(Z > 0) + \mathbf{P}_2(Z > 0) \geqslant 1$, puisque $\mathbf{P}_2(Z > 0) =$
$$= \int\limits_{f_2(x)>0} f_2(x)\mu^n(dx) = 1. \blacktriangleleft$$

---

*) Il est clair que, si $\varphi(c)$ est continue en c_ϵ, la résolution de (3) se ramène à la recherche du quantile d'ordre $1 - \epsilon$ de la distribution de Z.

Nous avons vu encore une fois que l'introduction des tests bayésiens randomisés a pour but d'assurer une variation « continue » des paramètres de ces tests (les valeurs possibles des risques des tests $\pi_{c,p}$ recouvrent l'intervalle]0, 1[tout entier). L'absence d'une telle variation continue des paramètres, liée au fait que sur un ensemble de $\mathbf{P}_1$-probabilité strictement positive est possible l'égalité $f_1(x) = cf_2(x)$, constitue le principal obstacle à la recherche de tests d'un niveau donné ou de tests minimax dans la classe des tests non randomisés. Cette situation prévaut aussi dans le cas d'un grand nombre d'hypothèses.

Il est important de noter également que deux types de tests optimaux — les tests les plus puissants et les tests minimax — sont bayésiens pour telle ou telle distribution *a priori*. Il est aisé de constater aussi que la classe des tests les plus puissants est confondue dans un certain sens avec celle des tests bayésiens. Cette situation dans laquelle l'approche bayésienne sert de base au choix des tests optimaux prévaudra dans la suite.

EXEMPLE 1. Considérons l'exemple 2 de l'Introduction. Les hypothèses H_1 et H_2 sont de la forme $H_1 = \{\mathrm{x}_i \in F(x)\}$, $H_2 = \{\mathrm{x}_i \in F(x - a)\}$, où $F(x)$ est une fonction de répartition donnée, a, un nombre donné. Supposons que $F(x)$ admet $f(x)$ pour densité et que la distribution de la quantité aléatoire $f(\mathrm{x}_1 - a)/f(\mathrm{x}_1)$ est continue. Le lemme de Neyman-Pearson (proposition 2 du théorème 1) nous dit que de tous les tests de niveau $1 - \epsilon$, le test

$$\prod_{i=1}^{n} \frac{f(\mathrm{x}_i - a)}{f(\mathrm{x}_i)} \geqslant c_\epsilon$$

sera le plus puissant pour éprouver l'hypothèse H_1 (l'objet est absent) contre l'hypothèse H_2 (l'objet est présent). Le nombre c_ϵ se détermine à partir de la condition

$$\mathbf{P}_1\left(\sum_{i=1}^{n} \ln \frac{f(\mathrm{x}_i - a)}{f(\mathrm{x}_i)} > \ln c_\epsilon\right) = \epsilon.$$

Pour les grands n on peut de toute évidence se servir du théorème limite central pour calculer cette probabilité.

§ 3*. Deux approches asymptotiques de calcul des tests. Comparaison numérique

1. Remarques préliminaires. Aux §§ 1 et 2 nous avons trouvé la forme des tests optimaux entre hypothèses simples. Le terme « calcul des tests » désignera le calcul des paramètres caractérisant le test. Dans le problème du

test le plus puissant pour $r = 2$, il s'agit de trouver les quantités c et p pour $\epsilon > 0$ donné et de déterminer le risque de deuxième espèce $\alpha_2(\pi_{c_\epsilon, p})$ ou, ce qui est équivalent, la puissance du test, $\beta(\pi_{c_\epsilon, p}) = 1 - \alpha_2(\pi_{c_\epsilon, p})$. On peut envisager ce problème sous un angle différent. Nous avons vu que pour $r =$ $= 2$ tous les tests optimaux sont de la forme des fonctions $\pi_{c, p}$ représentées dans (2.1). Soit donné un test $\pi_{c, p}$. Comment déterminer ses risques $\alpha_i(\pi_{c, p})$?

Cette question se pose aussi pour $r > 2$ pour le test (1.7), mais dans ce paragraphe on se limitera, par souci de simplicité, seulement au cas de deux hypothèses simples.

On développe plus bas les approches asymptotiques qui permettent de résoudre *approximativement* (pour de grands n) ces problèmes. Des approches analogues peuvent être utilisées pour le calcul des tests envisagés dans la suite.

Soit donné un test (2.1). Supposons pour simplifier que la distribution de $Z(X)$ est continue, de sorte que nous pouvons poser $p = 1$. Le test (2.1) (qui sera désigné par δ_c) devient alors non randomisé, et il nous faut calculer

$$\alpha_1(\delta_c) = \mathbf{P}_1\left(\frac{f_2(X)}{f_1(X)} \geqslant c\right), \tag{1}$$

$$\alpha_2(\delta_c) = \mathbf{P}_2\left(\frac{f_2(X)}{f_1(X)} < c\right).$$

Puisque $f_j(X) = \prod_{i=1}^{n} f_j(x_i)$, l'événement figurant sous le signe de la probabilité dans (1) peut être mis sous la forme

$$\sum_{i=1}^{n} \ln\frac{f_2(x_i)}{f_1(x_i)} \geqslant \ln c,$$

où les termes

$$\eta_i = \ln\frac{f_2(x_i)}{f_1(x_i)},$$

sont visiblement des variables aléatoires indépendantes équidistribuées dans chacun des cas $X \in \mathbf{P}_j$, $j = 1, 2$.

Le problème se ramène donc à l'étude des distributions des sommes $\sum_{i=1}^{n} \eta_i$ des variables aléatoires η_i.

On admettra dans la suite que la taille n de l'échantillon X croît indéfiniment. Ceci étant, par test on entendra en fait une suite de tests définis pour chaque n (nous avons utilisé cette convention pour les estimateurs dans le chapitre 2).

2. Hypothèses fixes. Nous admettrons dans ce numéro que les distributions $\mathbf{P}_i$ sont fixes, c'est-à-dire ne dépendent pas de la taille $n \to \infty$ de l'échantillon $X_n = [X_\infty]_n$. Soit à calculer un test le plus puissant de niveau fixé $1 - \epsilon$. On a

$$\mathsf{E}_1 \eta_i \equiv -a = \int f_1(x) \ln \frac{f_2(x)}{f_1(x)}\, \mu(dx) = -\rho_1(\mathbf{P}_1, \mathbf{P}_2) < 0,$$

$$\mathsf{E}_2 \eta_i \equiv b = \int f_2(x) \ln \frac{f_2(x)}{f_1(x)}\, \mu(dx) = \rho_1(\mathbf{P}_2, \mathbf{P}_1) > 0,$$

où ρ_1 est la distance de Kullback-Leibler (cf. § 2.21).

Ceci exprime en vertu de la loi des grands nombres que la $\mathbf{P}_1$-distribution de $\frac{1}{n} \sum_{i=1}^{n} \eta_i$ sera concentrée au voisinage du point $-a$, et la $\mathbf{P}_2$-distribution, au voisinage du point b. Et cette « distinction » des distributions sera la meilleure au sens du lemme de Neyman-Pearson. Posons $\sigma_j^2 = \mathbf{V}_j \eta_1$ et supposons que $\sigma_j^2 < \infty$. Alors

$$\alpha_1(\delta_c) = \mathbf{P}_1 \left(\sum_{i=1}^{n} \eta_i \geqslant \ln c \right) =$$

$$= \mathbf{P}_1 \left(\frac{1}{\sigma_1 \sqrt{n}} \sum_{i=1}^{n} (\eta_i + a) \geqslant \frac{\ln c + an}{\sigma_1 \sqrt{n}} \right). \tag{2}$$

Prenons pour $c = c(n)$ une suite quelconque telle que

$$\frac{\ln c + an}{\sigma_1 \sqrt{n}} \to \lambda_\epsilon,$$

où λ_ϵ est comme toujours le quantile d'ordre $1 - \epsilon$ de la distribution normale. De (2) et du théorème limite central il vient alors que

$$\alpha_1(\delta_c) \sim 1 - \Phi \left(\frac{\ln c + an}{\sigma_1 \sqrt{n}} \right) \to \epsilon. \tag{3}$$

DÉFINITION 1. Le test π qui vérifie la relation

$$\lim_{n\to\infty} \alpha_1(\pi) = \lim_{n\to\infty} \mathbf{E}_1\pi(X) = \epsilon$$

s'appelle *test de niveau asymptotique* $1 - \epsilon$.

Donc, pour

$$\ln c = -an + \lambda_3\sigma_1\sqrt{n} + o(\sqrt{n}) \tag{4}$$

le test δ_c sera de niveau asymptotique $1 - \epsilon$.

La relation (4) peut être traitée comme la solution approchée de l'équation pour un nombre c_ϵ tel que $\alpha_1(\delta_{c_\epsilon}) = \epsilon$.

Posons pour fixer les idées $\ln c = -an + \lambda_\epsilon\sigma_1\sqrt{n}$ et déterminons pour la valeur c choisie le comportement asymptotique du risque de deuxième espèce

$$\alpha_2(\delta_c) = \mathbf{P}_2\left(\sum_{i=1}^{n} \eta_i < \ln c\right) = \mathbf{P}_2\left(\sum_{i=1}^{n} \eta_i < -an + \lambda_\epsilon\sigma_1\sqrt{n}\right) =$$

$$= \mathbf{P}_2\left(\frac{1}{\sigma_2\sqrt{n}} \sum_{i=1}^{n} (\eta_i - b) < -\frac{(a+b)\sqrt{n}}{\sigma_2} + \frac{\lambda_\epsilon\sigma_1}{\sigma_2}\right). \tag{5}$$

Vu que $-\dfrac{(a+b)\sqrt{n}}{\sigma_2} + \dfrac{\lambda_\epsilon\sigma_1}{\sigma_2} \to -\infty$ lorsque $n \to \infty$, le théorème limite central nous dit seulement que $\alpha_2(\delta_c) \to 0$.

Le calcul du comportement asymptotique exact du second membre de (5) *nous conduit au calcul des probabilités des grands écarts des sommes de variables aléatoires* η_j.

Citons les résultats relatifs aux probabilités des grands écarts, développés dans le § 5 du chap. 7 [11]. Soit à calculer le comportement asymptotique de $\mathbf{P}\left(\sum_{i=1}^{n} \xi_i > x\right)$ lorsque $n \to \infty$ et $x \to \infty$, où ξ_i sont des variables indépendantes équidistribuées. Supposons que la distribution de ξ_i admet une composante absolument continue et que

$$\psi(\lambda) = \mathbf{E}e^{\lambda\xi} < \infty$$

pour certains $\lambda > 0$. Supposons par ailleurs que

$$\begin{aligned}\lambda_+ &= \sup\{\lambda : \psi(\lambda) < \infty\},\\ \Lambda(\alpha) &= -\inf_{\lambda}\{-\alpha\lambda + \ln\psi(\lambda)\}\end{aligned} \tag{6}$$

et que $\lambda(\alpha)$ est la valeur de λ qui réalise $\inf\{\cdot\}$.

On a alors la proposition suivante. (Cf. théorèmes 9, 10 du § 5 du chap. 7 [11]. Les conditions $\mathbf{V}\xi_i = 1$, $\mathbf{E}\xi_i = 0$ de ces théorèmes n'ont pas d'importance.)

THÉORÈME 1. *Supposons que* $\frac{x - n\mathbf{E}\xi_1}{\sqrt{n}} \to \infty$, *de sorte que*

$$\limsup_{n\to\infty} \frac{x}{n} < \alpha_+ = \frac{\psi'(\lambda_+)}{\psi(\lambda_+)}.$$

Alors l'équation pour le point $\lambda(\alpha)$

$$\alpha\psi(\lambda) = \psi'(\lambda) \tag{7}$$

admet une solution unique pour $\alpha < \alpha_+$ *et*

$$\mathbf{P}\left(\sum_{i=1}^{n} \xi_i > x\right) \sim \frac{1}{\sigma(\alpha)|\lambda(\alpha)|\sqrt{2\pi n}} \exp\{-n\Lambda(\alpha)\}, \tag{8}$$

où

$$\alpha = \frac{x}{n}, \quad \sigma^2(\alpha) = \frac{\psi''(\lambda(\alpha))}{\psi(\lambda(\alpha))} - \alpha^2.$$

Par ailleurs,

$$\Lambda(\mathbf{E}\xi_1) = 0, \quad \Lambda'(\alpha) = \lambda(\alpha),$$

$$\Lambda''(\alpha) = \lambda'(\alpha) = \frac{\psi(\lambda(\alpha))}{\psi''(\lambda(\alpha)) - \alpha^2\psi(\lambda(\alpha))}.$$

Revenons maintenant au calcul du comportement asymptotique de la quantité $\alpha_2(\delta_c)$ définie dans (5) et qui est égale à

$$\mathbf{P}_2\left(-\sum_{i=1}^{n} \eta_i > an - y\sqrt{n}\right) = \mathbf{P}_2\left(\sum_{i=1}^{n} (-\eta_i + b) > (a+b)n - y\sqrt{n}\right)$$

pour $y = \lambda_\epsilon \sigma_1$. Pour pouvoir appliquer le théorème cité, il faut poser

$$\xi_i = -\eta_i = \ln\frac{f_1(x_i)}{f_2(x_i)}, \quad x = an - y\sqrt{n}.$$

On obtient alors pour $0 \leqslant \lambda \leqslant 1$

$$\psi(\lambda) = \mathbf{E}_2 e^{-\lambda\eta_1} = \int f_2(x)(f_1(x)/f_2(x))^\lambda \mu(dx) =$$
$$= \int f_1^\lambda(x) f_2^{1-\lambda}(x)\mu(dx) \leqslant \left(\int f_1(x)\mu(dx)\right)^\lambda \left(\int f_2(x)\mu(dx)\right)^{1-\lambda} = 1.$$

Il s'ensuit également que $\psi(\lambda)$ sera finie dans un voisinage du point $\lambda = 1$ si

$$\int f_1(x)(f_1(x)/f_2(x))^\gamma \mu(dx) < \infty \tag{9}$$

pour un $\gamma > 0$. D'autre part, l'équation pour le point $\lambda(\alpha)$ sera de la forme

$$-\alpha + \frac{\psi'(\lambda)}{\psi(\lambda)} = 0,$$

ou

$$\psi'(\lambda) = \int f_2(x)(f_1(x)/f_2(x))^\gamma \ln \frac{f_1(x)}{f_2(x)} \mu(dx) = $$
$$= \alpha \int f_2(x)(f_1(x)/f_2(x))^\lambda \mu(dx). \tag{10}$$

Si $\alpha = a = \rho_1(\mathbf{P}_1, \mathbf{P}_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} \mu(dx)$, alors (10) sera vérifiée pour $\lambda = 1$. Ceci exprime que

$$\lambda(a) = 1, \quad \psi(\lambda(a)) = \psi(1) = 1.$$

D'où il vient

$$\Lambda(a) = a\lambda(a) - \ln \psi(\lambda(a)) = a,$$

$$\psi''(\lambda(a)) = \psi''(1) = \int f_1(x) \left(\ln \frac{f_1(x)}{f_2(x)}\right)^2 \mu(dx),$$

$$\sigma^2(a) = \psi''(1) - a^2 = \sigma_1^2,$$
$$\Lambda'(a) = \lambda(a) = 1, \quad \Lambda''(a) = \sigma_1^{-2}.$$

Les conditions du théorème 1 seront réunies si

1) la $\mathbf{P}_2$-distribution de $\ln \frac{f_1(\mathrm{x}_1)}{f_2(\mathrm{x}_1)}$ admet une composante absolument continue ;

2) $\int f_1(x)(f_1(x)/f_2(x))^\gamma \mu(dx) < \infty$ pour un $\gamma > 0$.

Vu que dans notre cas les fonctions $\sigma(\alpha)$, $\lambda(\alpha)$, $\Lambda''(\alpha)$ sont continues au voisinage du point $\alpha = a$ et que $\alpha = x/n = a - y/\sqrt{n}$, on trouve que $\Lambda(\alpha) = a - \frac{y}{\sqrt{n}} + \frac{y^2}{2\sigma_1^2 n} + o\left(\frac{1}{n}\right)$.

On peut désormais formuler le corollaire suivant du théorème cité.

Corollaire 1. *Supposons que la condition* (9) *est remplie et que la* $\mathbf{P}_2$*-distribution de* $\ln \frac{f_1(\mathrm{x}_1)}{f_2(\mathrm{x}_1)}$ *admet une composante absolument continue.*

Pour $n \to \infty$, on a alors

$$\alpha_2(\delta_c) = \mathbf{P}_2\Big(-\sum_{i=1}^{n} \eta_i > an - y\sqrt{n}\Big) \sim$$

$$\sim \frac{1}{\sigma_1\sqrt{2\pi n}} \exp\{-na + y\sqrt{n} - y^2/(2\sigma_1^2)\} =$$

$$= \frac{1}{\sigma_1\sqrt{2\pi n}} \exp\{-n\rho_1(\mathbf{P}_1, \mathbf{P}_2) + \lambda_c\sigma_1\sqrt{n} - \lambda_c^2/2\}. \quad (11)$$

Donc, $\alpha_2(\delta_c)$ décroît exponentiellement *) lorsque $n \to \infty$.

Il est immédiat de voir que si l'on fixe c dans (1), les deux probabilités $\alpha_1(\delta_c)$ et $\alpha_2(\delta_c)$ décroîtront exponentiellement, de même que $\alpha_Q(\delta_Q)$ pour toute Q fixe.

Comme

$$\mathbf{E}_1 e^{\lambda\eta_1} = \int f_1(x)\left(\frac{f_2(x)}{f_1(x)}\right)^{\lambda} \mu(dx) = \psi(1 - \lambda),$$

$$\min_{\lambda} \psi(\lambda) = \min_{\lambda} \psi(1 - \lambda),$$

les risques $\alpha_1(\delta_c)$ et $\alpha_2(\delta_c)$ décroîtront avec la même vitesse (la dépendance par rapport à n sera la même). Ceci exprime que le test minimax sera associé à un certain c fixe dont on peut trouver sans peine une valeur approchée en résolvant l'équation $\alpha_1(\delta_c) = \alpha_2(\delta_c)$ et en effectuant une analyse asymptotique du second membre de (8) pour $\alpha = c/n$, $n \to \infty$.

---

*) Nous obtenons incidemment la possibilité de donner encore une définition de la distance de Kullback-Leibler

$$\rho_1(\mathbf{P}_1, \mathbf{P}_2) = -\lim_{n\to\infty} \frac{1}{n} \ln \alpha_2(\delta_c) = -\lim_{n\to\infty} \frac{1}{n} \inf_{\delta\in K_\epsilon} \ln \alpha_2(\delta).$$

Signalons à ce propos que la $\mathbf{P}_2$-probabilité que la fonction de répartition empirique F_n^* tombe dans un voisinage de la fonction de répartition F_1 de $\mathbf{P}_1$ admet le même ordre de petitesse que $\exp\{-n\rho_1(\mathbf{P}_1, \mathbf{P}_2)\}$. Plus exactement, si $\delta = \delta(n) \to 0$ assez lentement, on a

$$-\lim_{n\to\infty} \frac{1}{n} \ln \mathbf{P}_2(\sup_x |F_n^*(x) - F_1(x)| < \delta) = \rho_1(\mathbf{P}_1, \mathbf{P}_2) \quad (12)$$

(théorème de Sanov). Donc, la distance $\rho_1(\mathbf{P}_1, \mathbf{P}_2)$ revêt une importante signification probabiliste. Le lecteur peut établir la relation (12) à l'aide du théorème 6, § 2 chap. V dans [11], en surmontant des difficultés insignifiantes.

L'approximation exponentielle (11) agit suffisamment bien pour les grands n si seulement l'écart normé

$$\frac{x + n\mathbf{E}_2\eta}{\sigma_2\sqrt{n}} = \frac{\sqrt{n}}{\sigma_2}(\rho_1(\mathbf{P}_1, \mathbf{P}_2) + \rho_1(\mathbf{P}_2, \mathbf{P}_1)) - \frac{\lambda_\epsilon \sigma_1}{\sigma_2} \tag{13}$$

est assez grand (cf. énoncé du théorème).

Dans les problèmes d'application où le nombre n est de l'ordre de 100, cette condition est remplie assez rarement et la valeur (13) est souvent voisine de 1. Ceci complique l'usage de la méthode de calcul de $\alpha_2(\delta_c)$ décrite et correspond à la situation où $\alpha_2(\delta_c)$ et $\alpha_1(\delta_c)$ ne sont pas très petits (de l'ordre de 0,1 par exemple). Dans le même temps, des valeurs de n de l'ordre de 100 suffisent pour appliquer avec succès le théorème limite central dans la zone des « écarts normaux ».

Donc, le problème qui nous préoccupe est de savoir quand pouvons-nous nous servir des approximations normales

$$\begin{aligned} \alpha_1(\delta_c) &= \mathbf{P}_1\left(\sum_{i=1}^{n} \eta_i \geqslant \ln c\right) \approx 1 - \Phi\left(\frac{\ln c - n\mathbf{E}_1\eta_1}{\sigma_1\sqrt{n}}\right), \\ \alpha_2(\delta_c) &= \mathbf{P}_2\left(\sum_{i=1}^{n} \eta_i < \ln c\right) \approx \Phi\left(\frac{\ln c - n\mathbf{E}_2\eta_1}{\sigma_2\sqrt{n}}\right) \end{aligned} \tag{14}$$

pour calculer $\alpha_1(\delta_c)$ et $\alpha_2(\delta_c)$.

On peut établir les formules (14) par une autre méthode conjecturant la proximité des hypothèses H_1 et H_2.

3. Hypothèses voisines. On envisagera un échantillon X dans un schéma de séries et on admettra que les distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ dépendent de n de telle sorte que

$$\rho_1(\mathbf{P}_1, \mathbf{P}_2) + \rho_1(\mathbf{P}_2, \mathbf{P}_1) \to 0 \tag{15}$$

lorsque $n \to \infty$, et la suite (13) converge vers une limite finie strictement positive.

Pour alléger les raisonnements et les rendre utiles pour la suite, on se bornera ici au cas paramétrique où $X \in \mathbf{P}_\theta$,

$$H_1 = \{\theta = \theta_1\}, \quad H_2 = \{\theta = \theta_2\}$$

et la famille $\{\mathbf{P}_\theta\}$ satisfait les conditions de régularité (RR) (cf. § 2.24).

Faisons tout d'abord quelques remarques formelles pour éclairer le fond du problème. Nous envisageons des hypothèses voisines, c'est-à-dire que nous supposons que $\theta_2 = \theta_1 + \delta$, où δ est petit. Le logarithme du rapport de vraisemblance sur lequel est construit un test le plus puissant est de

la forme *)

$$\ln \frac{f_{\theta_2}(X)}{f_{\theta_1}(X)} \sim \delta L'(X, \theta_1). \tag{16}$$

La statistique $U = L'(X, \theta)$, la principale partie de (16), s'appelle parfois *contribution efficace*. Si l'hypothèse H_1 est vraie, on a

$$\mathbf{E}_{\theta_1} U = 0, \quad \mathbf{V}_{\theta_1} U = nI(\theta_1).$$

Comme

$$L'(X, \theta_1) - L'(X, \theta_2) \sim \delta L''(X, \theta_2), \quad \mathbf{E}_{\theta_2} L''(X, \theta_2) = -nI(\theta_2),$$

il vient

$$\begin{aligned} \mathbf{E}_{\theta_2} U &\sim \delta nI(\theta_2) \sim \delta nI(\theta_1), \\ \mathbf{V}_{\theta_2} U &\sim nI(\theta_2) \sim nI(\theta_1). \end{aligned}$$

Ce qui exprime que les distributions de U seront distinctes sous les hypothèses H_1 et H_2 et pour de grands n si seulement la quantité $\mathbf{E}_{\theta_2} U - \mathbf{E}_{\theta_1} U \sim \sim \delta nI(\theta_1)$ est sensiblement plus grande ou comparable à $\sqrt{\mathbf{V}_{\theta_i} U} \sim \sqrt{nI(\theta_1)}$. En d'autres termes, on doit avoir l'égalité $\delta n = v\sqrt{n}$, $v \neq 0$, ou, ce qui est équivalent, $\delta = v/\sqrt{n}$.

Passons à un exposé plus rigoureux et supposons que

$$\theta_2 = \theta_1 + v/\sqrt{n}, \tag{17}$$

où les quantités θ_1 et v seront supposées fixes.

Suivant les notations du chapitre 2 posons

$$Z_i(t) = \frac{f_{\theta_i + t}(X)}{f_{\theta_i}(X)}, \quad Y_i(v) = \ln Z_i\left(\frac{v}{\sqrt{n}}\right).$$

Alors

$$\sum_{i=1}^{n} \eta_i = \ln \frac{f_{\theta_2}(X)}{f_{\theta_1}(X)} = Y_1(v) = -Y_2(-v). \tag{18}$$

Le théorème 2.29.3 nous donne pour $X \in \mathbf{P}_{\theta_1}$

$$Y_1(v) = \xi_n v - \frac{1}{2} v^2 (I(\theta_1) + \epsilon_n), \tag{19}$$

*) Le symbole $\sim$ exprime ici l'équivalence asymptotique pour $\delta \to 0$.

où $\epsilon_n \underset{\mathbf{P}_{\theta_1}}{\rightarrow} 0$, $\xi_n I^{-1/2}(\theta_1) \in \Phi_{0,1}$. De façon analogue, pour $X \in \mathbf{P}_{\theta_2}$ on a

$$-Y_2(-v) = \xi_n v + \frac{1}{2} v^2(I(\theta_2) + \epsilon_n),$$

où $\epsilon_n \underset{\mathbf{P}_{\theta_2}}{\rightarrow} 0$, $\xi_n I^{-1/2}(\theta_2) \in \Phi_{0,1}$.

Comme $I(\theta_2) \rightarrow I(\theta_1)$, on trouve que si l'hypothèse H_j, $j = 1, 2$, est vraie,

$$\sum_{i=1}^{n} \eta_i \Rightarrow \xi |v| \sqrt{I(\theta_1)} + (-1)^j \frac{v^2}{2} I(\theta_1), \quad \xi \in \Phi_{0,1}.$$

Ceci exprime que le théorème 2.29.3 entraîne le

COROLLAIRE 2. *Soient remplies les conditions* (*RR*) *et* (17). *Pour tout c fixe, on a alors les formules* (14) *ou, de façon plus exacte,*

$$\alpha_1(\delta_c) = \mathbf{P}_{\theta_1}\left(\sum_{i=1}^{n} \eta_i \geqslant \ln c\right) \rightarrow 1 - \Phi\left(\frac{\frac{v^2}{2} I(\theta_1) + \ln c}{|v| \sqrt{I(\theta_1)}}\right),$$
$$\alpha_2(\delta_c) = \mathbf{P}_{\theta_2}\left(\sum_{i=1}^{n} \eta_i < \ln c\right) \rightarrow \Phi\left(\frac{-\frac{v^2}{2} I(\theta_1) + \ln c}{|v| \sqrt{I(\theta_1)}}\right). \tag{20}$$

DÉFINITION 2. Les tests π_1 et π_2 sont dits *asymptotiquement équivalents* si

$$\limsup_{n \rightarrow \infty} |\alpha_j(\pi_1) - \alpha_j(\pi_2)| = 0, \quad j = 1, 2.$$

Un test π s'appelle *asymptotiquement le plus puissant* s'il est asymptotiquement équivalent à un test le plus puissant.

Vu que $\xi_n = L'(X, \theta_1) n^{-1/2}$ dans les représentations (18) et (19), on déduit de ces dernières que le test δ de région critique

$$\frac{vL'(X, \theta_1)}{\sqrt{nI(\theta_1)}} > vd, \quad d = \frac{v^2 I(\theta) + 2 \ln c}{2|v| \sqrt{I(\theta_1)}}$$

(le signe de v est important ici), aura les mêmes valeurs limites $\alpha_i(\delta)$ que le test δ_c et par suite sera asymptotiquement le plus puissant.

Par ailleurs, en vertu des résultats du § 2.29,

$$\xi_n = L'(X, \theta_1)/\sqrt{n} = (\hat{\theta}^* - \theta_1)\sqrt{n} I(\theta_1)(1 + \epsilon_n(X, \theta_1)),$$

$\epsilon_n(X, \theta_1) \underset{P_{\theta_i}}{\to} 0$. D'où il vient que le test de région critique

$$v(\hat{\theta}^* - \theta_1)\sqrt{nI(\theta_1)} > vd \tag{21}$$

est aussi asymptotiquement le plus puissant.

Pour obtenir un test δ_c le plus puissant de niveau asymptotique $1 - \epsilon$, il suffit de poser $d = \lambda_\epsilon$ dans (20). Le risque de deuxième espèce $\alpha_2(\delta_c)$ converge vers $\Phi(-v\sqrt{I(\theta_1)} + \lambda_\epsilon)$.

Pour $c = 1$, les deux limites de (20) prennent la même valeur

$$\lim_{n\to\infty} \alpha_j(\delta_c) = \Phi(-v\sqrt{I(\theta_1)}/2).$$

Dans ce cas il est naturel d'appeler le test δ_c (comparer avec le théorème 1.2) *test asymptotiquement minimax*.

4. Comparaison des approches asymptotiques. Exemple numérique. Dans les numéros 2 et 3 nous avons examiné deux approches asymptotiques dont l'usage était justifié dans des conditions définies et qui permettent de déterminer les valeurs approchées des risques de première et de deuxième espèce d'un test le plus puissant *). Ces formules sont données par (3) et (11) pour des *hypothèses fixes* et par (14) et (20) pour des *hypothèses voisines*. Les formules (11) et (20) ont été acquises à l'aide de (8) et de (14). C'est pourquoi on accordera si possible la préférence à ces dernières.

Nous avons déjà signalé que si $\alpha_1(\delta)$ et $\alpha_2(\delta)$ étaient petits (de l'ordre de 0,01 et moins), il était plus payant d'utiliser l'approche liée aux hypothèses fixes. En effet, il importe d'avoir une approximation suffisamment bonne qui est assurée par les formules (8), mais pas par le théorème limite central. Si $\alpha_1(\delta)$ et $\alpha_2(\delta)$ sont de l'ordre de 0,1 (disons $\geqslant 0,1$), on peut recommander la deuxième approche, en considérant la deuxième hypothèse $H_2 = \{\theta = \theta_2\}$ comme un élément de la suite d'hypothèses voisines $H_{2,n} = \{\theta = \theta_1 + v/\sqrt{n}\}$, où il faut poser de toute évidence $v = \sqrt{n}(\theta_2 - \theta_1)$ pour θ_1 et θ_2 donnés. Vu que les valeurs prévisibles de $\alpha_1(\delta)$ et de $\alpha_2(\delta)$ ne sont pas très petites, la valeur absolue de $v/\sqrt{I(\theta_1)}$ ne doit pas être élevée.

EXEMPLE 1. Considérons maintenant un exemple numérique illustrant dans une certaine mesure le lien existant entre les deux méthodes d'approximation proposées ci-dessus.

---

*) Signalons que parallèlement aux deux approches proposées on peut envisager tout un spectre de cas intermédiaires que, dans le langage paramétrique, on peut représenter sous la forme (cf. (17)) $\theta_2 = \theta_1 + zn^{-\gamma}$, $0 \leqslant \gamma \leqslant 1/2$. Les hypothèses voisines de cette nature présentent de l'intérêt lorsqu'il faut choisir les formules d'approximation décrivant le mieux la situation étudiée.

Soit $X \in \mathbf{\Gamma}_{\theta,\,1}$, c'est-à-dire que la densité des x_i est

$$f_\theta(x) = \theta e^{-\theta x}, \quad x \geqslant 0,$$

et supposons que l'hypothèse de base H_1 est de la forme $H_1 = \{\theta = 1\}$. Les hypothèses alternatives seront les hypothèses simples $H_2^{(1)} = \{\theta = 0{,}5\}$, $H_2^{(2)} = \{\theta = 0{,}8\}$, $H_2^{(3)} = \{\theta = 0{,}9\}$.

On éprouve l'hypothèse H_1 contre l'une des hypothèses $H_2^{(j)}, j = 1, 2, 3$ au vu de l'échantillon X. Donc $\theta_1 = 1$, quant à θ_2 il peut prendre trois valeurs: 0,5; 0,8 et 0,9, dont les deux dernières seront traitées comme correspondant à des hypothèses voisines de H_1. Déterminons les tests pour des échantillons de taille $n = 30, 100, 300, 1000$.

On a

$$\eta_i = \ln \frac{f_{\theta_2}(x_i)}{f_{\theta_1}(x_i)} = \ln \theta_2 - (\theta_2 - 1)x_i, \tag{22}$$

$$l'(\theta_1, x_i) = 1 - x_i, \tag{23}$$

$$\hat{\theta}^* = 1/\bar{x}.$$

D'où il découle que le test δ_c le plus puissant, ainsi que les deux tests asymptotiquement les plus puissants envisagés plus haut (de régions critiques $\sum l'(x_i, \theta_1) < d_1$ et $\hat{\theta}^* - \theta_1 < d_1/(nI(\theta_1))$, $d_1 = d\sqrt{nI(\theta_1))}$ seront de la forme $\delta_c(X) = H_2^{(j)}$ si

$$\sum_{i=1}^{n} (x_i - 1) > d_1. \tag{24}$$

Si $X \in \mathbf{\Gamma}_{1,\,1}$ (hypothèse H_1), on a

$$\mathbf{E}_1 x_1 = 1, \mathbf{V}_1 x_1 = 1 = I(1) = \mathbf{E}_1[l'(x_1, 1)]^2.$$

Si donc l'on pose $d_1 = 2\sqrt{n}$, il vient (comparer avec (14))

$$\alpha_1(\delta_c) = \mathbf{P}_1\left(\sum_{i=1}^{n} (x_i - 1) > d_1\right) =$$

$$= \mathbf{P}_1\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - 1) > 2\right) \to 1 - \Phi(2) \approx 0{,}023 \tag{25}$$

lorsque $n \to \infty$. Puisque $\sum_{i=1}^{n} \eta_i = n \ln \theta_2 + (1 - \theta_2) \sum_{i=1}^{n} x_i$, on déduit que

dans (14) (ou dans (20)) ln c est relié à d_1 par la relation

$$\ln c = n(\ln \theta_2 + 1 - \theta_2) + (1 - \theta_2)d_1.$$

Nous citons plus bas trois tableaux dans chacun desquels d_1 est supposé choisi de telle sorte que soit réalisée (25) (c'est-à-dire que $d_1 = 2\sqrt{n}$). Dans le premier tableau, on compare les vraies valeurs de $\alpha_1(\delta_c)$ à l'approximation (25). Dans le deuxième, on donne les vraies valeurs de $\alpha_2(\delta_c)$ et les valeurs approchées calculées à l'aide des formules des grands écarts (8). Dans le troisième enfin, on compare les vraies valeurs de $\alpha_2(\delta_c)$ aux valeurs approchées obtenues par les formules des hypothèses voisines (14). A noter que l'on se sert des approximations (8) et (14) et non pas des deuxièmes approximations (11) et (20) qui sont entachées d'erreurs supplémentaires. Les calculs sont développés plus bas.

Les nombres des tableaux 1, 2 et 3 sont donnés au centième près.

Tableau 1. **Valeurs de $\alpha_1(\delta_c)$.**

Le premier rang représente les valeurs exactes, le deuxième, les valeurs approchées (14)

n	30	100	300	1000
	0,031 0,023	0,028 0,023	0,026 0,023	0,024 0,023

Tableau 2. **Valeurs de $\alpha_2(\delta_c)$.**

Le premier rang représente les vraies valeurs, le deuxième, les valeurs approchées (8) ou (26) (les grands écarts)

θ_2 \ n	30	100	300	1000
0,5	0,028 0,033	$15 \cdot 10^{-7}$ $15 \cdot 10^{-7}$	$19 \cdot 10^{-19}$ $19 \cdot 10^{-19}$	$18 \cdot 10^{-72}$ $18 \cdot 10^{-72}$
0,8	0,71 —	0,35 —	0,028 0,033	$33 \cdot 10^{-8}$ $34 \cdot 10^{-8}$
0,9	0,89 —	0,79 —	0,53 —	0,085 0,11

La comparaison des tableaux 2 et 3 montre qu'en vertu des remarques faites plus haut, l'approximation basée sur les grands écarts agit mieux dans la partie supérieure droite du tableau (où $(\theta_1 - \theta_2)\sqrt{n} = (1 - \theta_2)\sqrt{n} > 3$), et l'approximation basée sur les hypothèses voisines, dans la partie inférieure gauche (où $(1 - \theta_2)\sqrt{n} < 3$). Les « blancs » expriment que l'utilisation de la méthode considérée n'a pas de sens. Dans le tableau 2 par exemple,

l'approximation (8) ne passe pas lorsque $\alpha_2(\delta_c) > 0,1$. Le calcul de $\alpha_2(\delta_c)$, lorsqu'il est inférieur, disons, à 10^{-6}, présente rarement un intérêt pratique. Les très petites valeurs de $\alpha_2(\delta_c)$ dans le tableau 2 pour $\theta_2 = 0,5$ et $n = 300$ et 1000 ont été calculées uniquement pour comparer les résultats.

Tableau 3. **Valeurs de $\alpha_2(\delta_c)$.**

Le premier rang représente les vraies valeurs, le deuxième, les approximations (14) (les hypothèses voisines)

θ_2 \ n	30	100	300	1000
0,5	0,028 0,041	$15 \cdot 10^{-7}$ $31 \cdot 10^{-6}$	$19 \cdot 10^{-19}$ —	$18 \cdot 10^{-72}$ —
0,8	0,71 0,69	0,35 0,35	0,028 0,031	$33 \cdot 10^{-8}$ $12 \cdot 10^{-7}$
0,9	0,89 0,89	0,79 0,79	0,53 0,52	0,085 0,086

Pour achever la discussion de ces tableaux, il nous reste à dire comment ont été calculées les vraies valeurs de $\alpha_i(\delta_c)$, $i = 1, 2$, et en quoi se transforment les approximations (8) et (14) dans notre cas.

On a

$$\alpha_2(\delta_c) = \mathbf{P}_{\theta_2}\left(\sum_{i=1}^{n} (x_i - 1) < 2\sqrt{n}\right).$$

Comme $\mathbf{E}_{\theta_2} x_i = 1/\theta_2$, $\mathbf{V}_{\theta_2} x_i = 1/\theta_2^2$, l'approximation normale (14) de $\alpha_2(\delta_c)$ sera de la forme

$$\Phi\left(\frac{\theta_2}{\sqrt{n}}\left[\left(1 - \frac{1}{\theta_2}\right)n + 2\sqrt{n}\right]\right) = \Phi((\theta_2 - 1)\sqrt{n} + 2\theta_2).$$

Considérons maintenant la formule (8) dans laquelle il faut poser $\xi_i = x_i$, $x = -n - 2\sqrt{n}$. La condition du théorème 1

$$\frac{x - n\mathbf{E}\xi_1}{\sqrt{n}} = \frac{-n - 2\sqrt{n} + n/\theta_2}{\sqrt{n}} = \sqrt{n}\left(\frac{1 - \theta_2}{\theta_2}\right) - 2 \to \infty$$

est remplie. Par ailleurs,

$$\psi(\lambda) = \mathbf{E}_{\theta_2} e^{-\lambda x_i} = \theta_2 \int_0^\infty e^{-\lambda x - \theta_2 x} dx = \frac{\theta_2}{\lambda + \theta_2},$$

$$\lambda_+ = \infty, \quad \alpha_+ = \lim_{\lambda \to \infty} \frac{\psi'(\lambda)}{\psi(\lambda)} = 0,$$

$$\alpha = \frac{x}{n} = -1 - \frac{2}{\sqrt{n}}.$$

Puisque $\lim_{n \to \infty} \alpha = -1 < 0$, la condition $\lim_{n \to \infty}.\sup \frac{x}{n} < \alpha_+$ est également satisfaite. L'équation (7) s'écrit

$$\frac{\alpha\theta_2}{\lambda + \theta_2} = -\frac{\theta_2}{(\lambda + \theta_2)^2},$$

et sa solution est $\lambda(\alpha) = -1/\alpha - \theta_2$. D'où

$$\Lambda(\alpha) = -\ln(-\alpha\theta_2) - 1 - \alpha\theta_2, \quad \sigma^2(\alpha) = 1/\lambda'(\alpha) = \alpha^2.$$

En vertu de (8) on obtient donc

$$\alpha_2(\delta_c) = \mathbf{P}_{\theta_2}\left(\sum_{i=1}^{n} \xi_i > c\right) = \mathbf{P}_{\theta_2}\left(\sum_{i=1}^{n} (x_i - 1) < 2\sqrt{n}\right) \sim$$

$$\sim \frac{1}{(1 + \alpha\theta_2)\sqrt{2\pi n}} \exp\{n[\ln(-\alpha\theta_2) + 1 + \alpha\theta_2]\}. \qquad (26)$$

En admettant que $\alpha = -1 - 2/\sqrt{n}$, on obtient les formules qui ont servi à calculer les valeurs de $\alpha_2(\delta_c)$ dans le tableau 2 (deuxième rang).

Signalons pour la comparaison que le second membre de (11) se transforme en l'expression

$$\frac{1}{(1 - \theta_2)\sqrt{2\pi n}} \exp\{n[\ln\theta_2 + 1 - \theta_2] + 2(1 - \theta_2)\sqrt{n} - 2\} \qquad (27)$$

que l'on peut déduire de (26) en y posant $\alpha = -1 - 2/\sqrt{n}$ et en limitant le développement en série aux termes d'ordre $< 1/\sqrt{n}$.

Le premier facteur du dénominateur de (26) qui est égal à $\sigma(\alpha)|\lambda(\alpha)| = 1 + \alpha\theta_2 = 1 - \theta_2 - 2\theta_2/\sqrt{n}$ se transforme dans (27) en $\sigma_1 = 1 - \theta_2$. Si θ_2 est voisin de 1, l'erreur relative liée au terme correctif $-2\theta_2/\sqrt{n}$ peut être considérable. Par exemple, pour $\theta_2 = 0{,}8$ et $n = 100$, on obtient $2\theta_2/\sqrt{n} = 0{,}16$, $\sigma_1 = 1 - \theta_2 = 0{,}2$, $\sigma(\alpha)|\lambda(a)| = 0{,}2 - 0{,}16 = 0{,}04$, de sorte que le premier facteur de (27) est de 5 fois (!) supérieur à celui de (26). Cet exemple montre que lorsque les hypothèses sont voisines et le facteur σ_1 de (11), petit, il faut manipuler les approximations (11) (ou (27)) avec beaucoup de précautions.

Pour calculer les vraies valeurs de $\alpha_i(\delta_c)$, on s'est servi du fait suivant. Supposons que $\eta(t)$ est un processus de renouvellement (cf. [11]) pour une

promenade de sauts $x_1, x_2, \dots$:

$$\eta(t) = \min\left\{k : \sum_{i=1}^{k} x_i \geqslant t\right\}.$$

Si $x_i \in \boldsymbol{\Gamma}_{\theta,1}$, alors, comme indiqué au § 4 du chap. 13, [11], le processus $\xi(t) = \eta(t) - 1$ est pour $t > 0$ un processus de Poisson de paramètre θ, c'est-à-dire que

$$\mathbf{P}(\eta(t) - 1 = k) = e^{-\theta t}\frac{(\theta t)^k}{k!}.$$

Remarquons maintenant que $\left\{\sum_{i=1}^{n} x_i \geqslant t\right\} = \{\eta(t) \leqslant n\}$, donc

$$\mathbf{P}_\theta\left(\sum_{i=1}^{n} x_i \geqslant t\right) = \sum_{k=0}^{n-1} e^{-\theta t}\frac{(\theta t)^k}{k!}. \tag{28}$$

Pour $t = n + 2\sqrt{n}$, on a donc

$$\alpha_1(\delta_c) = \mathbf{P}_1\left(\sum_{i=1}^{n} x_i \geqslant t\right) = \sum_{k=0}^{n-1} e^{-t}\frac{t^k}{k!},$$

$$\alpha_2(\delta_c) = \mathbf{P}_{\theta_2}\left(\sum_{i=1}^{n} x_i < t\right) = 1 - \sum_{k=0}^{n-1} e^{-\theta_2 t}\frac{(\theta_2 t)^k}{k!}.$$

Ces relations ont été utilisées pour le calcul des vraies valeurs de $\alpha_i(\delta_c)$, $i = 1, 2$.

Signalons qu'en plus de (28) on peut établir d'autres formules pour la distribution de $\sum_{i=1}^{n} x_i$, basées sur le fait que $\sum_{i=1}^{n} x_i \in \boldsymbol{\Gamma}_{\theta,n}$.

5. Lien entre le test le plus puissant et l'efficacité asymptotique de l'estimateur du maximum de vraisemblance. En se servant des calculs et résultats des §§ 1, 2, on peut prouver maintenant le théorème 2.25.3 relatif à l'efficacité asymptotique de l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ dans la classe $\tilde{K}^0$ des estimateurs asymptotiquement centrés (l'appartenance de $\hat{\theta}^*$ à $\tilde{K}^0$ a été établie au n° 2.29.3).

DÉMONSTRATION du théorème 2.25.3. Supposons par absurde qu'il existe un estimateur asymptotiquement normal θ^* tel que pour un θ_1 quel-

conque

$$\lim_{n\to\infty} \mathsf{E}_{\theta_1} n(\theta^* - \theta_1)^2 = \sigma^2(\theta_1) < I^{-1}(\theta_1) = \lim_{n\to\infty} \mathsf{E}_{\theta_1} n(\hat{\theta}^* - \theta_1)^2.$$

Soit à éprouver l'hypothèse $H_1 = \{X \in \mathbf{P}_{\theta_1}\}$ contre $H_2 = \{X \in \mathbf{P}_\theta, \theta = \theta_1 + vn^{-1/2}\}$. Construisons à cet effet un test δ de la forme suivante:

$$\delta(X) = \begin{cases} H_1 & \text{si } \theta^* \leqslant \theta_1 + vn^{-1/2}, \\ H_2 & \text{sinon,} \end{cases}$$

où pour fixer les idées on conviendra que $v > 0$. Alors

$$\alpha_1(\delta) = \mathbf{P}_{\theta_1}(\theta^* > \theta_1 + vn^{-1/2}) =$$

$$= \mathbf{P}_{\theta_1}\left(\frac{(\theta^* - \theta_1)\sqrt{n}}{\sigma(\theta_1)} > \frac{v}{\sigma(\theta_1)}\right) \to 1 - \Phi\left(\frac{v}{\sigma(\theta_1)}\right).$$

L'appartenance de θ^* à $\tilde{K}^0$ exprime que

$$\alpha_2(\delta) = \mathbf{P}_\theta(\theta^* \leqslant \theta_1 + vn^{-1/2}) = \mathbf{P}_\theta(\theta^* \leqslant \theta) \to 1/2.$$

Considérons maintenant un autre test $\delta_0(X)$ de région critique

$$\hat{\theta}^* - \theta_1 > (v + \gamma)/\sqrt{n}, \quad \gamma > 0,$$

qui, comme déjà établi, sera asymptotiquement le plus puissant (cf. (21)). Vu que

$$(v + \gamma)\sqrt{I(\theta_1)} < v/\sigma(\theta_1)$$

pour $\gamma > 0$ assez petit, on a pour ce test

$$\lim_{n\to\infty} \alpha_1(\delta_0) = 1 - \Phi((v + \gamma)\sqrt{I(\theta_1)}) > 1 - \Phi(v/\sigma(\theta_1)),$$

$$\lim_{n\to\infty} \alpha_2(\delta_0) = \lim_{n\to\infty} \mathbf{P}_\theta(\hat{\theta}^* \leqslant \theta + \gamma/\sqrt{n}) > 1/2.$$

Ceci exprime qu'à partir d'un certain n le test δ sera meilleur qu'un test le plus puissant. Cette contradiction prouve le théorème. ◄

§ 4. Test de choix entre hypothèses multiples. Classes de tests optimaux

1. Position du problème et notions fondamentales. Aux §§ 1 et 2 nous avons examiné les problèmes les moins compliqués de test d'hypothèses simples. Mais les hypothèses éprouvées sont souvent de nature plus complexe. Dans le cas paramétrique par exemple, l'hypothèse peut être de la forme $\{X \in \mathbf{P}_\theta; \theta \in \Theta_1\}$, où Θ_1 est un sous-ensemble donné de Θ. Une telle hypothèse ne détermine pas la distribution de l'échantillon de façon unique.

Toute hypothèse H qui n'est pas simple est dite *multiple* ou *composée*. Exemple : les hypothèses $\{X \in \Phi_{0,\sigma^2}\,;\, \sigma \geqslant 0\}$, $\{X \in \Phi_{\alpha,1}\,;\, \alpha \geqslant 0\}$.

Dans toute la suite de ce chapitre, on considérera des problèmes de choix entre *deux* hypothèses que l'on désignera par H_1 et H_2. Dans les prochains paragraphes, on se bornera à étudier le *cas paramétrique* $X \in \mathbf{P}_\theta$, $\theta \in \Theta$. Les hypothèses peuvent alors s'écrire

$$H_i = \{X \in \mathbf{P}_\theta ; \theta \in \Theta_i\}, \quad \Theta_i \subset \Theta, \quad \Theta_1 \cap \Theta_2 = \varnothing .$$

Puisque les valeurs de θ qui n'appartiennent pas à $\Theta_1 \cup \Theta_2$ ne sont pas envisagées, on peut, sans nuire à la généralité, admettre que $\Theta = \Theta_1 \cup \Theta_2$ et que H_2 est l'hypothèse *complémentaire* (ou *contraire*) de H_1, de sorte que l'hypothèse H_2 peut être représentée également sous la forme $H_2 = \{H_1$ *est fausse*$\}$. Comme dans le § 2, une des hypothèses — ici H_1 — sera l'*hypothèse de base*. Les hypothèses simples $H_\theta = \{X \in \mathbf{P}_\theta\}$, $\theta \in \Theta_2$, seront dites *alternatives* ou *concurrentes*, ou encore *contre-hypothèses*.

Le choix de l'hypothèse de base caractérise souvent l'attitude du chercheur vis-à-vis de l'objet étudié. L'hypothèse de base correspond en principe à une certaine conception, l'hypothèse alternative, à des écarts par rapport à cette conception qu'il faut soit prouver, soit réfuter. En général, on a affaire à une ou quelques hypothèses de base et à une énorme quantité d'hypothèses alternatives.

La procédure d'acceptation des hypothèses repose sur les tests statistiques. Puisque nous envisageons deux hypothèses en tout, comme au § 2 tout test (randomisé) π sera défini de façon unique par une fonction mesurable $\pi(x)$, $0 \leqslant \pi(x) \leqslant 1$, qui détermine la probabilité $\pi(X)$ d'accepter l'hypothèse H_2 pour chaque échantillon X (le choix aléatoire avec la probabilité $\pi(X)$ doit être effectué à l'aide d'un dispositif auxiliaire). Comme au § 2, la fonction $\pi(x)$ est dite *critique*. Si δ est un test non randomisé, la fonction $\pi(x) = \delta(x)$ ne prend que les valeurs 0 et 1 ; la région Ω_2 de l'espace $\mathscr{X}^n$ dans laquelle $\delta(x) = 1$ (la région d'acceptation de l'hypothèse H_2) est dans ce cas dite *critique*. On l'identifie souvent au test δ.

DÉFINITION 1. On appelle *risque de première espèce* d'un test π le nombre

$$\alpha_1(\pi) = \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi(X).$$

Si le test n'est pas randomisé, il est évident que

$$\alpha_1(\delta) = \sup_{\theta \in \Theta_1} \mathbf{P}_\theta(X \in \Omega_2).$$

Ceci est la probabilité maximale (par rapport à $\theta \in \Theta_1$) de rejeter à tort l'hypothèse H_1. Pour faciliter la recherche des tests optimaux on considère

généralement des tests π vérifiant la condition

$$\alpha_1(\pi) = \varepsilon \quad (\text{ou } \alpha_1(\pi) \leqslant \epsilon).$$

La classe de ces tests sera désignée par K_ϵ.

Le nombre $1 - \alpha_1(\pi) = 1 - \epsilon$ sera appelé *niveau* (ou *seuil*) *de signification* *) du test π.

Du point de vue statistique, l'utilisation d'un test $\delta \in K_\epsilon$ exprime que dans une longue série d'expériences visant à éprouver une hypothèse H_1 à l'aide de δ, on se trompera dans moins de ϵ% des cas si l'hypothèse H_1 est vraie.

Le niveau de signification d'un test est arbitraire. Mais en règle générale on prend pour ϵ l'une des valeurs standards suivantes : 0,005 ; 0,01 ; 0,05 ; 0,1. Cette standardisation permet de réduire le volume des tables utilisées par le statisticien. Il n'existe aucune autre raison spéciale au choix de ces valeurs. Le choix du niveau de signification de π doit tenir compte de la *puissance* du test

$$\beta_\pi(\theta) = \mathbf{E}_\theta \pi(X), \quad \theta \in \Theta_2.$$

Si elle est trop faible, il faut probablement envisager une plus petite valeur du niveau $1 - \epsilon$.

Notre attitude vis-à-vis de l'hypothèse avant l'expérience est un élément important qui peut influencer le choix du niveau de signification. Si l'on est fermement convaincu de la véracité de l'hypothèse (la probabilité *a priori* $\mathbf{Q}(H_1)$ est élevée du point de vue bayésien), il faut des preuves irréfutables pour ébranler notre conviction. Dans ces conditions, il faut envisager des tests de niveau élevé et un ϵ très petit (il est alors très peu probable que l'on tombe dans la région critique Ω_2 si H_1 est vraie).

On s'en tiendra à la conception développée lors de la construction des intervalles de confiance, conception qui consiste en ce qui suit : si la probabilité ϵ d'un événement A est petite, on admet qu'il est pratiquement impossible que cet événement ait lieu en une seule expérience.

Certains statisticiens préconisent le point de vue suivant : il n'y a aucune raison de fixer le niveau d'un test et son choix n'est guidé par

*) Souvent c'est ϵ et non $1 - \epsilon$ qui est appelé niveau de signification. Mais ceci est un peu illogique : en effet il est plus naturel d'admettre que plus le niveau de signification est élevé et plus le test est « significatif ». C'est justement à partir de ces considérations que nous avons défini le seuil de signification (ou de confiance) pour les intervalles de confiance. Vu qu'il existe un lien direct (cf. § 8) entre les tests et les intervalles de confiance, il serait irraisonnable de modifier la terminologie pour étudier les tests.

aucune règle raisonnable. Ils traitent le test des hypothèses non point comme une procédure débouchant nécessairement sur l'acceptation d'une des deux hypothèses, mais comme un processus qui se déroule dans l'esprit du chercheur et qui définit l'attitude de ce dernier vis-à-vis des hypothèses. Dans cette optique, au niveau de signification fixé on préférera le niveau « réellement atteint » qui se définit comme suit. Considérons une famille de tests non randomisés δ de niveau $1 - \epsilon$, où $\epsilon \in]0, 1[$, et désignons par $\Omega_{2,\epsilon}$ la région critique de δ en admettant que $\Omega_{2,\epsilon_2} \subset \Omega_{2,\epsilon_1}$ pour $\epsilon_2 < \epsilon_1$.

DÉFINITION 2. On appellera *niveau réellement atteint* d'une famille de tests δ sur un échantillon X la variable aléatoire $1 - \epsilon(X)$, où

$$\epsilon(X) = \inf \{\epsilon : X \in \Omega_{2,\epsilon}\}.$$

Plus la quantité $1 - \epsilon(X)$ est élevée et plus l'hypothèse H_1 est contestée par l'échantillon X.

La valeur de $\epsilon(X)$ permet d'accepter ou de rejeter l'hypothèse pour tout niveau $1 - \epsilon$ donné à l'avance, par une simple comparaison de $\epsilon(X)$ à ϵ.

EXEMPLE 1. Dans le paragraphe précédent, nous avons construit un test le plus puissant de l'hypothèse $H_1 = \{X \in \mathbf{\Gamma}_{1,1}\}$ contre l'hypothèse $H_2 = \{X \in \mathbf{\Gamma}_{1/2,1}\}$. Ce test admet pour région critique

$$\Omega_2 = \left\{x \in \mathscr{X}^n : \sum_{i=1}^{n} (x_i - 1) > d_1\right\}.$$

Supposons que pour un échantillon X de taille $n = 10$, l'on ait $\sum_{i=1}^{10} x_i = 18$. Vu que dans le cas de l'hypothèse H_1 on a $\sum_{i=1}^{n} x_i \in \mathbf{\Gamma}_{1,n}$ et $\mathbf{\Gamma}_{1,n}(]a, b[) = \mathbf{H}_{2n}(]2a, 2b[)$, il vient $\mathbf{\Gamma}_{1,10}(]18, \infty[) = \mathbf{H}_{20}(]36, \infty[) = 0{,}0154$ (cf. table III ou [8]) et le niveau réellement atteint sera dans ce cas égal à $1 - \epsilon(X) = 1 - 0{,}0154 = 0{,}9846$, de sorte que l'hypothèse H_1 sera réfutée par un test le plus puissant de niveau $1 - \epsilon = 0{,}98$ et ne le sera pas par un test le plus puissant de niveau $1 - \epsilon = 0{,}99$.

2. Tests uniformément les plus puissants. Revenons aux tests rendomisés π que nous sommes convenus de définir par une fonction critique $\pi(x)$, $x \in \mathscr{X}^n$. (La fonction $\pi(x)$ peut être appelée aussi *fonction de décision statistique* (randomisée).)

S'il existe une statistique exhaustive $S(X)$, on peut se limiter à des tests $\pi(X)$ qui dépendent de X uniquement par l'intermédiaire de la statistique exhaustive $S(X)$, c'est-à-dire à des tests de la forme $\pi(X) = \varphi(S(X))$. On sait en effet que toute l'information sur le paramètre inconnu est concentrée dans S, et l'intervention d'autres statistiques (d'une autre information sur X) n'a pas de sens.

Nous avons déjà signalé que pour déterminer les tests optimaux, on restreint généralement l'ensemble des tests envisagés à la classe K_ϵ des tests de niveau fixé. Parmi ces tests on pourrait essayer de chercher celui dont la puissance

$$\beta_\pi(\theta) = \mathbf{E}_\theta \pi(X)$$

est maximale dans le domaine Θ_2 (autrement dit, dont le risque de deuxième espèce $1 - \beta_\pi(\theta)$ serait minimal). En d'autres termes, la probabilité d'accepter à juste titre l'hypothèse H_2 doit être maximale.

La fonction $\beta_\pi(\theta) = \mathbf{E}_\theta \pi(X)$ est souvent appelée aussi *puissance* du test π.

DÉFINITION 3. On dit qu'un test $\pi^\circ \in K_\epsilon$ est *uniformément le plus puissant* dans K_ϵ si pour tout $\pi \in K_\epsilon$ on a

$$\beta_{\pi^\circ}(\theta) \geqslant \beta_\pi(\theta), \quad \forall \theta \in \Theta_2. \tag{1}$$

Il est évident que les tests uniformément les plus puissants n'existent pas toujours. Si un tel test π° existe, le graphique de sa puissance $\beta_{\pi^\circ}(\theta)$ est situé au-dessus de celui de toute autre puissance $\beta_\pi(\theta)$ dans le domaine Θ_2 sous réserve qu'elles soient toutes deux $\leqslant \epsilon$ dans le domaine Θ_1 (on rappelle que $\alpha_1(\pi) = \sup_{\theta \in \Theta_1} \beta_\pi(\theta)$), de sorte que $\beta_{\pi^\circ}(\theta)$ est l'enveloppe de la famille $\{\beta_\pi(\theta)\}$ dans le domaine Θ_2.

Supposons que $\Theta_1 = \{\theta_1\}$, $\mathbf{E}_{\theta_1} \pi^\circ(X) = \epsilon$. Le test uniformément le plus puissant π° sera alors visiblement un test le plus puissant de niveau $1 - \epsilon$ entre l'hypothèse $\{\theta = \theta_1\}$ et son alternative $\{\theta = \theta_2\}$ pour tout $\theta_2 \in \Theta_2$. Vu que la forme du test le plus puissant est connue, on peut déterminer tout naturellement un test uniformément le plus puissant : on peut en effet trouver ce dernier si le test le plus puissant entre les hypothèses $\{\theta = \theta_1\}$ et $\{\theta = \theta_2\}$ est indépendant de θ_2.

La réciproque est vraie : si un test le plus puissant de K_ϵ entre les hypothèses $\{\theta = \theta_1\}$ et $\{\theta = \theta_2\}$, $\theta_2 \in \Theta_2$, dépend essentiellement de θ_2, c'est qu'il n'existe pas de test uniformément le plus puissant entre les hypothèses $\{\theta = \theta_1\}$ et $\{\theta \in \Theta_2\}$.

Si l'hypothèse H_2 est simple (Θ_2 est composé du seul point θ_2), la notion de test uniformément le plus puissant perd partiellement son sens et se transforme en un ordinaire test le plus puissant, c'est-à-dire en un test maximisant $\mathbf{E}_{\theta_2} \pi(X)$ dans la classe K_ϵ.

Définissons maintenant les tests bayésiens et minimax pour éprouver des hypothèses multiples.

3. Tests bayésiens. Pour tester les hypothèses multiples on se servira des deux approches bayésiennes suivantes.

a) *Approche totalement bayésienne*. Elle consiste à supposer que les hypothèses $H_\theta = \{X \in \mathbf{P}_\theta\}$, $\theta \in \Theta$, sont choisies au hasard avec une distribution *a priori* $\mathbf{Q}$. En d'autres termes, on définit une tribu $\mathfrak{C}$ sur $\Theta = \Theta_1 \cup \cup\ \Theta_2$, $\Theta_1 \in \mathfrak{C}$, $\Theta_2 \in \mathfrak{C}$, et on traite θ comme une variable aléatoire sur l'espace $(\Theta, \mathfrak{C}, \mathbf{Q})$.

La distribution $\mathbf{Q}$ induit des distributions $\mathbf{Q}_i$ sur Θ_i, $i = 1, 2$, et des probabilités $q_i = \mathbf{Q}(\theta \in \Theta_i)$, de sorte que $\mathbf{Q} = q_1\mathbf{Q}_1 + q_2\mathbf{Q}_2$. Désignons par $H_{\mathbf{Q}_i}$ l'hypothèse que $\theta \in \Theta_i$ est choisi au hasard avec la distribution $\mathbf{Q}_i$.

DÉFINITION 4. On dit qu'un test $\pi_{\mathbf{Q}}$ est *bayésien* si c'est un test bayésien entre deux hypothèses simples $H_{\mathbf{Q}_1}$ et $H_{\mathbf{Q}_2}$, associé à une distribution *a priori* (q_1, q_2) (cf. § 1).

b) *Approche partiellement bayésienne*. On admet ici que sont données des distributions *a priori* $\mathbf{Q}_i$ sur Θ_i mais que les probabilités *a priori* q_1 et q_2 sont inconnues. On a affaire alors à un test entre deux hypothèses simples $H_{\mathbf{Q}_1}$ et $H_{\mathbf{Q}_2}$.

Désignons comme précédemment

$$K_\epsilon = \{\pi : \sup_{\theta\in\Theta_1} \mathbf{E}_\theta \pi(X) \leqslant \epsilon\},$$

et posons

$$K_\epsilon^{\mathbf{Q}_1} = \{\pi : \mathbf{E}_{\mathbf{Q}_1} \pi(X) \leqslant \epsilon\},$$

où $\mathbf{E}_{\mathbf{Q}_i}$ représente l'espérance mathématique par rapport à la distribution engendrée sur $\Theta_i \times \mathscr{X}^n$ par $\mathbf{Q}_i$ et $\mathbf{P}_\theta$.

DÉFINITION 5. On dit qu'un test $\pi_{\mathbf{Q}_1, \mathbf{Q}_2}$ est *bayésien dans* $K_\epsilon^{\mathbf{Q}_1}$ si c'est un test le plus puissant de niveau $1 - \epsilon$ entre deux hypothèses simples $H_{\mathbf{Q}_1}$ et $H_{\mathbf{Q}_2}$.

Si l'une des hypothèses H_i dégénère en une hypothèse simple (Θ_1 ou Θ_2 est un singleton), il en sera de même de la distribution correspondante. Dans ce cas, nous simplifierons l'indice du test $\pi_{\mathbf{Q}_1, \mathbf{Q}_2}$ et écrirons par exemple $\pi_{\mathbf{Q}_1}$ au lieu de $\pi_{\mathbf{Q}_1, \mathbf{Q}_2}$ si $\Theta_2 = \{\theta_2\}$ est un singleton.

La construction des tests $\pi_{\mathbf{Q}_1, \mathbf{Q}_2}$ n'apporte aucune complication. Ces tests nous aideront à construire des tests uniformément les plus puissants et des tests minimax.

4. Tests minimax.

DÉFINITION 6. On dit qu'un test $\bar\pi$ entre les hypothèses $H_1 = \{\theta \in \Theta_1\}$ et $H_2 = \{\theta \in \Theta_2\}$ est *minimax* dans K_ϵ (resp. $K_\epsilon^{\mathbf{Q}_1}$) si $\bar\pi \in K_\epsilon$ (resp. $\bar\pi \in K_\epsilon^{\mathbf{Q}_1}$) et si est maximisé

$$\inf_{\theta\in\Theta_2} \mathbf{E}_\theta \pi(X) = \inf_{\theta\in\Theta_2} \beta_\pi(\theta).$$

Il serait plus correct d'appeler ce test, *test maximin* (c'est le minimum qui est maximisé). Mais nous opterons pour le terme « minimax » d'autant plus qu'il conserve son sens si l'on a affaire non pas à la puissance mais aux risques de deuxième espèce.

Les tests bayésiens et minimax seront étudiés plus en détail au § 9. Les prochains paragraphes seront consacrés à l'établissement des conditions autorisant la construction des tests uniformément les plus puissants.

§ 5. Tests uniformément les plus puissants

Dans ce paragraphe nous considérons deux importants cas particuliers mettant en jeu un paramètre scalaire θ et où il est possible de construire des tests uniformément les plus puissants. Nous obtiendrons de même un résultat utile relativement à la construction des tests les plus puissants.

1. Alternatives unilatérales. Rapport de vraisemblance monotone. Supposons que l'hypothèse de base est $H_1 = \{\theta \leqslant \theta_1\}$ et l'hypothèse alternative $H_2 = \{\theta > \theta_1\}$. Une telle hypothèse H_2 sera dite *unilatérale* contrairement, disons, à l'hypothèse $H_2 = \{\theta \neq \theta_1\}$ (complémentaire de $H_1 = \{\theta = \theta_1\}$) qui est bilatérale, car elle admet des écarts par rapport à θ_1 dans les deux sens.

L'autre condition posée consiste en ce qui suit. Supposons que la condition (A_0) est remplie et qu'il existe une fonction $T(x)$ telle que pour tous les $\theta, \theta_0, \theta > \theta_0$, le rapport de vraisemblance

$$\frac{f_\theta(x)}{f_{\theta_0}(x)} \tag{1}$$

est une fonction croissante (ou décroissante) de $T(x)$. On dit alors que la famille $\{\mathbf{P}_\theta\}$ possède un *rapport de vraisemblance monotone*.

La statistique T étant exhaustive, on a $f(x) = \psi(T(x), \theta)h(x)$ et la condition posée concernera le rapport $\psi(T, \theta)/\psi(T, \theta_0)$. Cette condition exprime que pour tous les $\theta > \theta_0$ et tout $d > 0$ l'inégalité $f_\theta(x)/f_{\theta_0}(x) \geqslant d$ peut être mise sous la forme $T(x) \geqslant c_n(\theta, \theta_0, d)$ (ou $T(x) \leqslant c_n(\theta, \theta_0, d)$).

Les familles $\{\Phi_{\alpha, 1}\}$ et $\{\Phi_{0, \sigma^2}\}$, par exemple, possèdent des rapports de vraisemblance monotones, puisque

$$\frac{f_{\alpha, 1}(X)}{f_{\alpha_0, 1}(X)} = \exp\left\{(\alpha - \alpha_0)n\bar{x} - \frac{n}{2}(\alpha^2 - \alpha_0^2)\right\},$$

$$\frac{f_{0, \sigma^2}(X)}{f_{0, \sigma_0^2}(X)} = \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2} - \frac{1}{\sigma_0^2}\right)\sum_{i=1}^{n} x_i^2\right\},$$

et les inégalités correspondantes seront de la forme ($\alpha > \alpha_0, \sigma > \sigma_0$)

$$\bar{x} \geqslant c_n(\alpha, \alpha_0, d) = \frac{1}{2} (\alpha + \alpha_0) + \frac{\ln d}{n(\alpha - \alpha_0)} \quad (T(X) = \bar{x}),$$

$$\sum_{i=1}^{n} x_i^2 \geqslant c_n(\sigma, \sigma_0, d) = \frac{2(\sigma\sigma_0)^2}{\sigma^2 - \sigma_0^2} \ln d \quad \left(T(X) = \sum_{i=1}^{n} x_i^2\right).$$

De nombreuses familles paramétriques du § 2.2 possèdent aussi un rapport de vraisemblance monotone. Dans la suite, on admettra pour fixer les idées que (1) est une fonction $T(x)$ *croissante*.

THÉORÈME 1. *Supposons que θ est un paramètre scalaire et que la famille $\{\mathbf{P}_\theta\}$ possède un rapport de vraisemblance monotone. Alors :*

1) *Dans K_ϵ il existe un test uniformément le plus puissant de $H_1 = \{\theta \leqslant \theta_1\}$ contre $H_2 = \{\theta > \theta_1\}$, qui est de la forme*

$$\pi^\circ(X) = \begin{cases} 1 & si \quad T(X) > c, \\ p & si \quad T(X) = c, \\ 0 & si \quad T(X) < c, \end{cases} \tag{2}$$

où c et p se déterminent à partir de la condition

$$\mathbf{E}_{\theta_1} \pi^\circ(X) = \mathbf{P}_{\theta_1}(T(X) > c) + p\mathbf{P}_{\theta_1}(T(X) = c) = \epsilon. \tag{3}$$

2) *La puissance $\beta^\circ(\theta) = \mathbf{E}_\theta \pi^\circ(X)$ est strictement croissante par rapport à θ pour tous les θ tels que $\beta^\circ(\theta) < 1$.*

3) *Pour tous les θ_0, le test (2) est uniformément le plus puissant entre les hypothèses $H_1^\circ = \{\theta \leqslant \theta_0\}$ et $H_2^\circ = \{\theta > \theta_0\}$ dans la classe $K_{\beta^\circ(\theta_0)}$.*

4) *Pour tout $\theta < \theta_1$, le test (2) minimise $\beta(\theta) = \mathbf{E}_\theta \pi(X)$ dans la classe K_ϵ.*

DÉMONSTRATION. Considérons tout d'abord les hypothèses simples $\{\theta = \theta_1\}$ et $\{\theta = \theta_2\}$, $\theta_2 > \theta_1$. Un test le plus puissant entre ces hypothèses dans la classe des tests π tels que $\mathbf{E}_{\theta_1} \pi(X) = \epsilon$ est, en vertu du théorème 2.1, de la forme (2), puisque l'inégalité $Z(X) > d$ est équivalente à $T(X) > c$ (moyennant une relation convenable entre c et d), où les constantes c et p se déterminent à partir de (3) (comparer avec (2.3)). Vu que les nombres c et p se déduisent de façon unique d'une équation de la forme (3), il vient que le test (2) sera aussi un test le plus puissant entre les hypothèses $\{\theta = \theta_0\}$ et $\{\theta = \theta_2\}$, $\theta_2 > \theta_0$ dans la classe $K_{\beta^\circ(\theta_0)}$. De là et du théorème 2.1 (cf. (2.4)), il s'ensuit que $\beta^\circ(\theta_2) > \beta^\circ(\theta_0)$.

Comme $\beta^\circ(\theta)$ est croissante, on a

$$\mathbf{E}_\theta \pi^\circ(X) \leqslant \epsilon \quad \text{pour} \quad \theta \leqslant \theta_1. \tag{4}$$

La classe K_ϵ des tests π vérifiant (4) est contenue dans la classe $\{\pi : \mathbf{E}_{\theta_1}\pi(X) = \epsilon\}$. Etant donné que le test (2) maximise $\beta(\theta_2)$ dans cette dernière classe, il le maximisera aussi dans K_ϵ. Il reste à remarquer que le test (2) est indépendant de θ_2, et par suite, les conclusions établies sont valables pour tout $\theta_2 > \theta_1$. Ceci prouve les trois premières assertions du théorème.

La quatrième proposition résulte des trois premières si on les applique pour éprouver l'hypothèse $H_1' = \{\theta \geqslant \theta_1\}$ contre l'hypothèse $H_2' = \{\theta < < \theta_1\}$ à l'aide d'un test uniformément le plus puissant de classe $\{\Pi(X) : \mathbf{E}_\theta\Pi(X) \leqslant 1 - \epsilon, \theta \geqslant \theta_1\}$ qui sera de la forme $\Pi^\circ(X) = 1 - \pi^\circ(X)$ et dont la puissance $1 - \beta^\circ(\theta) = \mathbf{E}_\theta\Pi^\circ(X)$ sera maximale pour $\theta < \theta_1$. ◄

Une importante classe de familles de distributions à rapport de vraisemblance monotone est la famille exponentielle à un paramètre (cf. § 2.15) dont la densité $f_\theta(x)$ est de la forme

$$f_\theta(x) = h(x) \exp\{a(\theta)U(x) + V(\theta)\}. \tag{5}$$

En effet, dans ce cas

$$\frac{f_\theta(x)}{f_{\theta_0}(x)} = \exp\left\{(a(\theta) - a(\theta_0)) \sum_{i=1}^{n} U(x_i) + n(V(\theta) - V(\theta_0))\right\},$$

et le rapport de vraisemblance dépendra monotonement de $T(x) = = \sum_{i=1}^{n} U(x_i)$ si $a(\theta) - a(\theta_0)$ conserve son signe pour tous les $\theta, \theta_0, \theta > \theta_0$.

COROLLAIRE 1. *Supposons que $f_\theta(x)$ est de la forme* (5), *où $a(\theta)$ est une fonction monotone. Il existe alors un test uniformément le plus puissant π° dans la classe K_ϵ entre les hypothèses $H_1 = \{\theta \leqslant \theta_1\}$ et $H_2 = \{\theta > \theta_1\}$. Si $a(\theta)$ est strictement croissante, ce test est de la forme* (2), (3). *Si elle est strictement décroissante, les inégalités dans* (2), (3) *changent de sens.*

A noter que si l'on éprouve une alternative bilatérale, par exemple l'hypothèse $H_1 = \{\theta = \theta_1\}$ contre l'hypothèse $H_2 = \{\theta \neq \theta_1\}$, il n'existe plus de test uniformément le plus puissant pour la famille exponentielle (5). En effet, supposons pour simplifier que $a(\theta)$ est strictement croissante et que la $\mathbf{P}_\theta$-distribution de $T(X) = \sum_{i=1}^{n} U(x_i)$ est absolument continue pour tous les θ. En vertu du théorème 2.1, le test le plus puissant entre $\{\theta = \theta_1\}$ et $\{\theta = \theta_2\}$ ne sera pas randomisé et admettra pour région critique le domaine $T(X) \geqslant c$ si $\theta_2 > \theta_1$, et le domaine $T(X) < c$ si $\theta_2 < \theta_1$. On voit que la puissance maximale au point θ_2 est réalisée sur des tests fondamentalement différents sèlon le signe de la différence $\theta_2 - \theta_1$. Du théorème 1 il s'ensuit

que si l'on prend l'un quelconque de ces tests, par exemple celui pour lequel $\pi(X) = 1$ dans la région $T(X) \geqslant c$, il sera uniformément le plus puissant pour tout $\theta_2 > \theta_1$ et visiblement pas pour $\theta_2 < \theta_1$.

Comme déjà signalé, les deux hypothèses simples du théorème 2.1 relatif au test le plus puissant sont dans un certain sens symétriques (le test le plus puissant minimise le risque de deuxième espèce $\alpha_2(\pi)$ si $\alpha_1(\pi)$ est fixe, et inversement minimise $\alpha_1(\pi)$ si $\alpha_2(\pi)$ est fixe). Cette symétrie fait défaut dans le test des hypothèses multiples. A cette circonstance est lié le fait intéressant suivant. Nous venons juste de voir que la famille exponentielle n'admet pas de test uniformément le plus puissant de $H_1 = \{\theta = \theta_1\}$ contre $H_2 = \{\theta \neq \theta_1\}$. Il est clair, de ce qui précède, qu'il n'existe pas non plus de test uniformément le plus puissant de l'hypothèse $\{\theta \in]\theta_1, \theta_2[\}$ contre l'hypothèse $\{\theta \notin]\theta_1, \theta_2[\}$. Mais si pour hypothèse de base H_1 on prend l'hypothèse $H_1 = \{\theta \notin]\theta_1, \theta_2[\}$ et pour son alternative l'hypothèse $H_2 = \{\theta \in]\theta_1, \theta_2[\}$, il existera alors un test uniformément le plus puissant dans la classe K_ϵ. Considérons maintenant le deuxième cas où il est possible de construire un test uniformément le plus puissant.

2. Hypothèse de base bilatérale. Famille exponentielle.

THÉORÈME 2. *Supposons que $f_\theta(x)$ est définie par* (5) *et soit à éprouver l'hypothèse $H_1 = \{\theta \notin]\theta_1, \theta_2[\}$, $\theta_1 < \theta_2$, contre l'hypothèse $H_2 = \{\theta \in]\theta_1, \theta_2[\}$. Si la fonction $a(\theta)$ est monotone, alors :*

1) *Dans la classe $K_\epsilon = \{\pi : \sup_{\theta \notin]\theta_1, \theta_2[} \mathbf{E}_\theta \pi(X) \leqslant \epsilon\}$ il existe un test uniformément le plus puissant π° de la forme*

$$\pi^\circ(x) = \begin{cases} 1 & si \quad T(x) \in]c_1, c_2[, \\ p_i & si \quad T(x) = c_i, i = 1, 2, \\ 0 & si \quad T(x) \notin [c_1, c_2], \end{cases} \tag{6}$$

où $T(x) = \sum_{i=1}^{n} U(x_i)$, et les constantes c_i et p_i se déterminent à partir des conditions

$$\mathbf{E}_{\theta_1} \pi^\circ(X) = \mathbf{E}_{\theta_2} \pi^\circ(X) = \epsilon. \tag{7}$$

2) *Ce test maximise la puissance $\beta(\theta) = \mathbf{E}_\theta \pi(X)$ sous la condition* (7) *à l'intérieur de l'intervalle $]\theta_1, \theta_2[$ et la minimise à l'extérieur* (cf. fig. 4).

3) *Pour $0 < \epsilon < 1$ la fonction $\beta^\circ(\theta)$ présente un maximum en un point $\theta_0 \in]\theta_1, \theta_2[$ et décroît strictement lorsque θ s'éloigne de θ_0 vers la droite ou la gauche. Ceci étant, nous excluons le cas où la distribution de $T(X)$ est concentrée en deux points, c'est-à-dire le cas où il existe des t_1 et t_2 tels que*

$$\mathbf{P}_\theta(T(X) = t_1) + \mathbf{P}_\theta(T(X) = t_2) = 1 \quad \textit{pour tous les } \theta. \tag{8}$$

La proposition suivante nous sera utile dans notre exposé.

LEMME 1. *Les équations* (7) *admettent toujours une solution en c_i et p_i, $i = 1, 2$, pour* $0 < \epsilon < 1$.

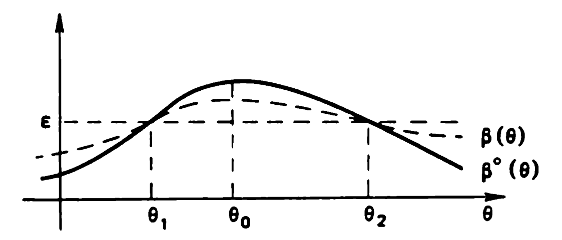


Fig. 4. Courbes de puissance $\beta^\circ(\theta) = \mathbf{E}_\theta\, \pi^\circ(X)$ et $\beta(\theta) = \mathbf{E}_\theta\, \pi(X)$ pour un test arbitraire $\pi \in K_\epsilon$.

La démonstration de ce lemme sera produite plus loin.

DÉMONSTRATION du théorème 2. Mettons la fonction de vraisemblance sous la forme

$$f_\theta(x) = c(\theta)e^{a(\theta)T(x)}h(x), \tag{9}$$

où nous admettrons, pour fixer les idées, que $a(\theta)$ est strictement croissante.

Considérons la position bayésienne suivante du problème. Soit à éprouver une hypothèse de base « mixte » H qui consiste en ce que $\{\theta = \theta_1\}$ avec la probabilité q et $\{\theta = \theta_2\}$ avec la probabilité $1 - q$ contre l'hypothèse $H_0 = \{\theta = \theta_0\}$, $\theta_0 \in]\theta_1, \theta_2[$. Supposons par ailleurs que les probabilités *a priori* des hypothèses H et H_0 sont respectivement égales à r et $1 - r$. Vu que H et H_0 définissent complètement la distribution de l'échantillon, on peut les traiter comme des hypothèses simples et appliquer les résultats du § 2. Un test bayésien (noté π°) sera alors de la forme

$$\pi^\circ(X) = \begin{cases} 1 & \text{si} \quad R(X) \equiv \dfrac{f_{\theta_0}(X)}{qf_{\theta_1}(X) + (1-q)f_{\theta_2}(X)} > \dfrac{r}{1-r}, \\ p(X) & \text{si} \quad R(X) = \dfrac{r}{1-r}, \\ 0 & \text{si} \quad R(X) < \dfrac{r}{1-r}. \end{cases} \tag{10}$$

En vertu de (9), l'inégalité $R(X) > r/(1-r)$ est équivalente à

$$q\,\frac{c(\theta_1)}{c(\theta_0)}\,e^{(a(\theta_1)-a(\theta_0))T} + (1-q)\,\frac{c(\theta_2)}{c(\theta_0)}\,e^{(a(\theta_2)-a(\theta_0))T} < \frac{1-r}{r}. \tag{11}$$

Le premier membre est une fonction convexe de T, puisque $a(\theta_1) - a(\theta_0) < 0$, $a(\theta_2) - a(\theta_0) > 0$. Ceci exprime que (11) peut être mise sous la forme

$$c_1 < T < c_2,$$

où $c_i = c_i(q, r)$; les nombres $c_1 < c_2$ prennent toutes les valeurs possibles. Posons la fonction $p(X)$ de (10) égale à p_1 si $T(X) = c_1$ et à p_2 si $T(X) = c_2$.

En vertu du lemme 1, il existe des nombres c_i, $i = 1, 2$, (ou ce qui est équivalent, des nombres q et r) et p_i tels que (7) soit remplie. Montrons maintenant que la fonction $\pi^\circ(X)$ définie dans (10) ou ce qui revient au même dans (6) sera douée de toutes les propriétés énumérées dans le théorème 2. Ce qui vient d'être dit exprime que nous avons commencé à traiter π° en même temps comme une fonction de décision pour éprouver H_1 contre H_2. Le test π° étant un test bayésien (de H contre H_0), pour tout autre test π on a

$$r[q\mathbf{E}_{\theta_1}\pi^\circ + (1 - q)\mathbf{E}_{\theta_2}\pi^\circ] + (1 - r)\mathbf{E}_{\theta_0}(1 - \pi^\circ) \leqslant \\ \leqslant r[q\mathbf{E}_{\theta_1}\pi + (1 - q)\mathbf{E}_{\theta_2}\pi] + (1 - r)\mathbf{E}_{\theta_0}(1 - \pi). \qquad (12)$$

Donc, si le test π vérifie comme π° la relation (7), alors

$$\mathbf{E}_{\theta_0}\pi^\circ \geqslant \mathbf{E}_{\theta_0}\pi.$$

Ceci exprime qu'en chaque point $\theta_0 \in]\theta_1, \theta_2[$, le test π° maximise la puissance $\beta(\theta) = \mathbf{E}_\theta\pi$ dans la classe des tests π vérifiant (7). Mais la condition (7) définit une classe de tests plus vaste que K_ϵ. Donc, π° maximisera $\beta(\theta)$ dans K_ϵ aussi. Etant indépendant de θ_0, le test π° sera uniformément le plus puissant dans K_ϵ.

Signalons encore qu'en vertu du théorème 2.1.

$$\beta^\circ(\theta_0) = \mathbf{E}_{\theta_0}\pi^\circ \geqslant \epsilon$$

et l'égalité n'est possible ici que lorsque

$$qf_{\theta_1}(x) + (1 - q)f_{\theta_2}(x) = f_{\theta_0}(x) \qquad (13)$$

$[\mu^n]$-presque partout.

De façon analogue, on s'assure à l'aide de (12) que π° minimisera $\mathbf{E}_{\theta_1}\pi$ pour $\mathbf{E}_{\theta_0}\pi$ et $\mathbf{E}_{\theta_2}\pi$ fixes (nous utilisons les mêmes raisonnements que pour la démonstration des théorèmes du § 1).

Montrons maintenant que π° minimise $\beta(\theta)$ à l'extérieur de $]\theta_1, \theta_2[$. Supposons que $\theta^\circ < \theta_1$. Remplaçons dans ce qui précède le triplet de points $(\theta_1, \theta_0, \theta_2)$ par le triplet $(\theta^\circ, \theta_1, \theta_2)$ et remarquons que pour le nouveau problème le test π° sera encore bayésien (en effet, sa forme ne dépend

pas du choix des points θ_i, $i = 0, 1, 2$) dans la classe des tests π tels que $\mathbf{E}_{\theta^\circ}\pi = \beta^\circ(\theta^\circ)$, $\mathbf{E}_{\theta_2}\pi = \epsilon$. Mais d'après la remarque faite ci-dessus le test π° minimisera $\mathbf{E}_{\theta^\circ}\pi$ pour $\mathbf{E}_{\theta_1}\pi$ et $\mathbf{E}_{\theta_2}\pi$ fixes. Ceci prouve les deux premières assertions du théorème.

Prouvons la troisième. Remarquons préalablement qu'un changement des variables d'intégration nous permet d'écrire

$$\mathbf{P}_\theta(T \in A) = c(\theta) \int_{\{x:\, T(x) \in A\}} e^{a(\theta)T(x)} h(x)\mu^n(dx) = c(\theta) \int_{t \in A} e^{a(\theta)t} \nu(dt),$$

où la mesure ν est définie par la relation

$$\nu(A) = \int_{\{x:\, T(x) \in A\}} h(x)\mu^n(dx).$$

Ceci exprime que la distribution de T par rapport à la mesure ν admet une densité (cf. également le lemme 2.15.1) $g_\theta(t) = c(\theta)e^{a(\theta)t}$ et par suite appartient aussi à une famille exponentielle. La fonction $a(\theta)$ étant monotone, on peut introduire un nouveau paramètre $\beta = a(\theta)$ sans rien modifier au problème et à ses hypothèses. Nous pouvons donc admettre sans restreindre la généralité que $a(\theta) = \theta$. Dans ce cas les fonctions $c(\theta) = \left[\int e^{\theta t}\nu(dt)\right]^{-1}$ et $\beta^\circ(\theta) = \mathbf{E}_\theta \pi^\circ(X)$ seront visiblement continues. Supposons maintenant que la proposition du théorème relative au caractère du comportement de $\beta^\circ(\theta)$ est fausse. Il existe alors trois points $\theta' < \theta'' < \theta'''$ tels que

$$\beta^\circ(\theta') = \beta^\circ(\theta'') = \beta^\circ(\theta''') = \alpha \in \,]0, 1[. \tag{14}$$

Nous avons vu que π° maximise $\beta(\theta'')$ sous réserve que $\beta(\theta') = \beta(\theta'') = \alpha$, ceci étant si la condition (13) n'est pas remplie, on aura $\beta^\circ(\theta'') > \alpha$. Mais dans notre cas la relation (13) exprime que

$$q\frac{f_{\theta'}}{f_{\theta''}} + (1-q)\frac{f_{\theta'''}}{f_{\theta''}} \equiv q\frac{c(\theta')}{c(\theta'')}e^{(\theta'-\theta'')T} + (1-q)\frac{c(\theta''')}{c(\theta'')}e^{(\theta'''-\theta'')T} = 1$$

ν-presque partout. Le premier membre étant convexe par rapport à T, cette égalité n'est possible que pour deux valeurs de T au plus. Donc, si (8) est exclue, $\beta^\circ(\theta'') > \beta^\circ(\theta') = \alpha$ et (14) est impossible. ◄

DÉMONSTRATION du lemme 1. Nous effectuerons la démonstration sous la condition simplificatrice que la distribution de $T(X)$ est continue, c'est-à-dire que $\mathbf{P}_\theta(T = c) = 0$ pour tous les θ et c. Ceci nous évitera des complications insignifiantes. Dans ce cas, les remarques faites en fin de

démonstration du théorème 2 nous permettent d'écrire

$$\mathbf{E}_\theta \pi^\circ(X) = \mathbf{P}_\theta(T \in]c_1, c_2[) = \int_{c_1}^{c_2} g_\theta(t)\nu(dt) = c(\theta) \int_{c_1}^{c_2} e^{\theta t}\nu(dt).$$

Cette fonction est continue par rapport à θ, c_1 et c_2.

Désignons par c_+ la valeur de c pour laquelle $\mathbf{P}_{\theta_1}(T \leqslant c_+) = 1 - \epsilon$. Sur $]-\infty, c_+[$ est alors définie une fonction $d(c)$ telle que

$$\mathbf{P}_{\theta_1}(T \in]c, d(c)[) = \int_c^{d(c)} g_{\theta_1}(t)\nu(dt) = \epsilon.$$

Il est clair que $d(c)$ est une fonction continue strictement croissante.

On démontrera la proposition annoncée lorsqu'on aura établi que la fonction

$$\psi(c) = \mathbf{P}_{\theta_2}(T \in]c, d(c)[) = \int_c^{d(c)} g_{\theta_2}(t)\nu(dt)$$

est continue et strictement croissante, $\psi(-\infty) < \epsilon$ et $\psi(c_+) > \epsilon$. Il existera alors un c_0 tel que $\psi(c_0) = \epsilon$ et par suite $\mathbf{P}_{\theta_i}(c_0, d(c_0)) = \epsilon, i = 1, 2$.

La continuité de $\psi(c)$ est évidente. Prouvons la monotonie. Mettons $\psi(c)$ sous la forme

$$\psi(c) = \int_c^{d(c)} g_{\theta_1}(t)r(t)\nu(dt), \tag{15}$$

où $r(t)$ est la densité de la $\mathbf{P}_{\theta_2}$-distribution de T par rapport à la $\mathbf{P}_{\theta_1}$-distribution :

$$r(t) = \frac{c(\theta_2)}{c(\theta_1)} e^{(\theta_2 - \theta_1)t}.$$

Supposons que Δ est, pour fixer les idées, tel que $c + \Delta < d(c)$. Comme

$$\int_c^{c+\Delta} g_{\theta_1}(t)\nu(dt) = \int_{d(c)}^{d(c+\Delta)} g_{\theta_1}(t)\nu(dt), \tag{16}$$

il vient alors

$$\psi(c + \Delta) - \psi(c) = \int_{d(c)}^{d(c+\Delta)} g_{\theta_1}(t)r(t)\nu(dt) - \int_c^{c+\Delta} g_{\theta_1}(t)r(t)\nu(dt) \geqslant$$

$$\geqslant [r(d(c)) - r(c + \Delta)]\lambda \geqslant 0,$$

où λ est la valeur commune de l'intégrale (16).

Assurons-nous maintenant que $\psi(-\infty) < \epsilon$. Désignons par t_0 la solution de l'équation $r(t) = 1$. Si $d(-\infty) \leqslant t_0$, alors $r(t) < 1$ sur l'intervalle $]-\infty, d(-\infty)[$, et l'inégalité annoncée est évidente en vertu de (15). Si en revanche $d(-\infty) > t_0$, on a de façon analogue

$$\psi(-\infty) = 1 - \mathbf{P}_{\theta_2}(T \in]d(-\infty), \infty[) <$$
$$< 1 - \mathbf{P}_{\theta_1}(T \in]d(-\infty), \infty[) = \mathbf{P}_{\theta_1}(T \in]-\infty, d(-\infty)[) = \epsilon.$$

On établit de la même façon que $\psi(c_+) > \epsilon$. ◄

REMARQUE 1. Nous laissons au lecteur le soin de s'assurer que pour $\theta_1 < \theta_2$ le théorème 2 et tout ce qui a été dit reste en vigueur si l'on remplace l'intervalle ouvert $]\theta_1, \theta_2[$ par l'intervalle fermé $[\theta_1, \theta_2]$, c'est-à-dire si l'on éprouve l'hypothèse $H_1 = \{\theta \notin [\theta_1, \theta_2]\}$ contre l'hypothèse $H_2 = \{\theta \in [\theta_1, \theta_2]\}$.

REMARQUE 2. De la démonstration du théorème il ressort que la condition d'exponentialité de la famille $\{\mathbf{P}_\theta\}$ peut être remplacée par la condition plus faible de convexité du rapport

$$q\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} + (1-q)\frac{f_{\theta_2}(X)}{f_{\theta_0}(X)}$$

par rapport à une statistique T (comparer avec (10), (11)).

REMARQUE 3. Attirons une fois de plus l'attention sur le fait que si l'hypothèse de base était $H_2 = \{\theta \in]\theta_1, \theta_2[\}$, il n'existerait pas de test uniformément le plus puissant, puisque dans ce cas les tests « unilatéraux » de la forme $T > c$ ou $T < c$ associés respectivement aux alternatives $\theta > \theta_2$ et $\theta < \theta_1$ seraient plus puissants qu'un test de la forme $T \notin]c_1, c_2[$. Pour les alternatives $\theta > \theta_2$ par exemple, il existera un test uniformément le plus puissant de la forme $T > c$, et la condition $\pi \in K_\epsilon$ nous conduit à une seule contrainte $\mathbf{E}_{\theta_2}\pi \leqslant \epsilon$ (cf. remarques de la fin du n° 2).

Néanmoins, si la classe K_ϵ est restreinte de façon naturelle (cf. §§ 6, 7), il existera un test uniformément le plus puissant dans ce problème aussi.

3. Autre approche des problèmes envisagés. La teneur mathématique de la principale proposition du théorème 2 ainsi que des théorèmes des §§ 1, 2 est très simple et mérite d'être mise en évidence. Dans le théorème 2 par exemple, elle consiste en le problème de calcul aux variations suivant : dans la classe des fonctions π vérifiant les conditions

$$\int \pi(x) f_{\theta_i}(x)\mu^n(dx) = \epsilon, \quad i = 1, 2,$$

trouver la fonction π° qui maximise

$$\int \pi(x) f_{\theta_0}(x)\mu^n(dx).$$

La proposition suivante est une *généralisation du lemme fondamental de Neyman-Pearson*.

LEMME 2. *Soient $f_1, \ldots, f_{m+1}$ des fonctions réelles définies sur $\mathscr{X}^n$ et intégrables par rapport à une mesure μ^n. Soient π des fonctions critiques telles que*

$$\int \pi(x) f_i(x) \mu^n(dx) = \epsilon_i, \quad i = 1, \ldots, m. \tag{17}$$

Alors l'élément π° qui maximise $\int \pi(x) f_{m+1}(x) \mu^n(dx)$ est de la forme

$$\pi^\circ(x) = \begin{cases} 1 & si \quad f_{m+1}(x) > \sum_{i=1}^{m} k_i f_i(x), \\ 0 & si \quad f_{m+1}(x) < \sum_{i=1}^{m} k_i f_i(x), \end{cases}$$

où $k_1, \ldots, k_m$ sont déterminés à partir des conditions (17).

DÉMONSTRATION. Désignons $F_i(\pi) = \int \pi(x) f_i(x) \mu^n(dx)$, $i = 1, \ldots, m + 1$. L'élément π vérifiant les conditions $F_i(\pi) = \epsilon_i$, $i = 1, \ldots, m$, maximise $F_{m+1}(\pi)$ si et seulement s'il maximise $F_{m+1}(\pi) - \sum_{i=1}^{m} k_i F_i(\pi)$ pour des $k_1, \ldots, k_m$ (la valeur de la somme est fixe ici). Il suffit donc que π maximise

$$\int \left(f_{m+1}(x) - \sum_{i=1}^{m} k_i f_i(x)\right) \pi(x) \mu^n(dx).$$

Or cette expression devient maximale si l'on pose $\pi(x) = 1$ là où l'expression $f_{m+1}(x) - \sum_{i=1}^{m} k_i f_i(x) > 0$, et $\pi(x) = 0$ là où elle est < 0. Les constantes k_i dont dépend π, et les valeurs « libres » de π sur l'ensemble $\left\{f_{m+1}(x) = \sum_{i=1}^{m} k_i f_i(x)\right\}$ doivent être choisies de façon à ce que (17) ait lieu. ◄

4. Approche bayésienne et distributions *a priori* les plus défavorables à la construction de tests les plus puissants et de tests uniformément les plus puissants. Le lemme 2 explique la teneur mathématique des constructions effectuées dans ce paragraphe. Dans ce numéro on abordera les mêmes

choses, mais sous un angle différent. En effet, en démontrant le théorème 2 nous avons implicitement utilisé une approche liée à la construction de tests minimax à l'aide de tests bayésiens (comparer avec le théorème 1.2). Cette approche sera étudiée plus en détail dans la suite. On se propose d'établir une proposition générale utile pour la construction des tests uniformément les plus puissants dans le cas général et de mettre en évidence son lien avec l'approche minimax.

Soit à éprouver l'hypothèse de base $H_1 = \{\theta \in \Theta_1\}$ contre l'hypothèse simple $H_2 = \{\theta = \theta_2\}$, $\theta_2 \notin \Theta_1$. Pour H_2 on peut prendre une contre-hypothèse quelconque $\{X \in \mathbf{G}\}$, où $\mathbf{G}$ admet une densité g par rapport à μ et n'est pas liée à la famille $\{\mathbf{P}_\theta\}$. Le problème consiste à chercher un test le plus puissant π de niveau $1 - \epsilon$ entre H_1 et H_2. En d'autres termes, il faut trouver une fonction π de K_ϵ

$$K_\epsilon = \left\{\pi : \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi(X) \leqslant \epsilon\right\}, \tag{18}$$

minimisant $\beta(\theta_2) = \mathbf{E}_{\theta_2}\pi(X)$. Dans les considérations précédentes nous avons plus d'une fois constaté une certaine « dualité » dans la position du problème : la maximisation de la puissance, à risque de première espèce fixe, équivaut à la minimisation du risque de première espèce, à puissance fixe. En inversant ainsi le problème, on est conduit à minimiser (18), c'est-à-dire à construire un test minimax (cette construction sera discutée plus en détail au § 9). Ceci explique dans une certaine mesure la similitude de la proposition prouvée plus bas avec le théorème 1.2.

Considérons donc la position partiellement bayésienne du problème, position dans laquelle $\theta \in \Theta_1$ est un paramètre aléatoire de distribution $\mathbf{Q}_1$. L'hypothèse multiple H_1 est alors remplacée par l'hypothèse simple $H_{\mathbf{Q}_1}$ sous laquelle la densité de X est définie comme la moyenne par rapport à la mesure $\mathbf{Q}_1$

$$f_{\mathbf{Q}_1}(x) = \int_{\Theta_1} f_\theta(x)\mathbf{Q}_1(d\theta).$$

Pour éprouver $H_{\mathbf{Q}_1}$ contre H_2 dans la classe $K_\epsilon^{\mathbf{Q}_1} = \{\pi : \mathbf{E}_{\mathbf{Q}_1}\pi(X) \leqslant \epsilon\}$ des tests de niveau $1 - \epsilon$, il existe un test le plus puissant $\pi_{\mathbf{Q}_1}$ de la forme ($\pi_{\mathbf{Q}_1}$ est le test $\pi_{\mathbf{Q}_1\mathbf{Q}_2}$ dans les notations du § 4, où $\mathbf{Q}_2$ est dégénérée au point θ_2) :

$$\pi_{\mathbf{Q}_1}(x) = \begin{cases} 1 & \text{si} \quad g(x) > cf_{\mathbf{Q}_1}(x), \\ 0 & \text{si} \quad g(x) < cf_{\mathbf{Q}_1}(x) \end{cases} \tag{19}$$

(ici $g(x) = f_{\theta_2}(x)$ dans le cas paramétrique).

THÉORÈME 3. *Supposons qu'il existe une distribution* $\mathbf{Q}_1$ *concentrée sur le sous-ensemble* $\Theta_1^\circ \subset \Theta_1$ $(\mathbf{Q}_1(\Theta_1^\circ) = 1)$, *telle que*

$$1) \qquad \pi_{\mathbf{Q}_1} \in K_\epsilon^{\mathbf{Q}_1}, \tag{20}$$

$$2) \qquad \mathbf{E}_\theta \pi_{\mathbf{Q}_1}(X) = \text{const} = \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi_{\mathbf{Q}_1}(X) \tag{21}$$

pour tous les $\theta \in \Theta_1^\circ$.

Alors le test $\pi_{\mathbf{Q}_1} \in K_\epsilon$ *est le plus puissant entre* H_1 *et* H_2.

DÉMONSTRATION. Assurons-nous tout d'abord que $\pi_{\mathbf{Q}_1} \in K_\epsilon$. D'après les hypothèses du théorème

$$\sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi_{\mathbf{Q}_1}(X) = \int_{\Theta_1^\circ} \mathbf{E}_\theta \pi_{\mathbf{Q}_1}(X) \mathbf{Q}_1(d\theta) = \mathbf{E}_{\mathbf{Q}_1} \pi_{\mathbf{Q}_1}(X) \leqslant \epsilon. \tag{22}$$

Supposons maintenant que π est un autre test quelconque de K_ϵ, c'est-à-dire un test de niveau $1 - \epsilon$ de H_1 contre H_2. Alors

$$\mathbf{E}_{\mathbf{Q}_1} \pi(X) = \int \pi(x) f_{\mathbf{Q}_1}(x) \mu^n(dx) = \int_{\Theta_1} \mathbf{E}_\theta \pi(X) \mathbf{Q}_1(d\theta) \leqslant \epsilon,$$

et par suite $\pi \in K_\epsilon^{\mathbf{Q}_1}$. On a donc en vertu de la définition de $\pi_{\mathbf{Q}_1}$

$$\mathbf{E}_{\theta_2} \pi_{\mathbf{Q}_1}(X) \geqslant \mathbf{E}_{\theta_2} \pi(X). \blacktriangleleft$$

La distribution $\mathbf{Q}_1$ du théorème est dite *la plus défavorable*. Ceci est lié à la circonstance suivante. La quantité $\beta_{\mathbf{Q}_1}(\theta_2) = \mathbf{E}_{\theta_2} \pi_{\mathbf{Q}_1}(X)$ est la plus grande valeur prise par la puissance sur la classe $K_\epsilon^{\mathbf{Q}_1}$ pour la distribution « *a priori* » $\mathbf{Q}_1$ sur Θ_1. Si l'on considère maintenant une autre distribution $\mathbf{Q}^{\cdot}$ sur Θ_1, on obtient

$$\beta_{\mathbf{Q}^{\cdot}}(\theta_2) \geqslant \beta_{\mathbf{Q}_1}(\theta_2), \qquad \beta_{\mathbf{Q}_1}(\theta_2) = \inf_{\mathbf{Q}^{\cdot}} \beta_{\mathbf{Q}^{\cdot}}(\theta_2)$$

(ceci est la signification du terme « la distribution la plus défavorable »). En effet, en vertu de (22), le test $\pi_{\mathbf{Q}_1}$ est de classe K_ϵ et par suite de classe $K_\epsilon^{\mathbf{Q}^{\cdot}}$. Ceci exprime que sa puissance $\beta_{\mathbf{Q}_1}(\theta_2) = \mathbf{E}_{\theta_2} \pi_{\mathbf{Q}_1}(X)$ sera au plus égale à celle d'un test le plus puissant dans $K_\varepsilon^{\mathbf{Q}^{\cdot}}$, puissance égale par définition à $\beta_{\mathbf{Q}^{\cdot}}(\theta_2)$.

Nous aurions pu prouver maintenant les théorèmes 1 et 2 à l'aide du théorème 3. L'ensemble Θ_1° sur lequel est concentrée la distribution la plus défavorable est composé dans les théorèmes 1 et 2 respectivement d'un seul point $\{\theta_1\}$ et de deux points $\{\theta_1, \theta_2\}$. Les conditions (20) et (21) se transforment respectivement en les conditions (3) et (7).

On se servira de façon analogue du théorème 3 pour construire un test uniformément le plus puissant dans les autres cas : si le test $\pi_{\mathbf{Q}_1}$ construit ne dépend pas de $\theta_2 \in \Theta_2$, il sera uniformément le plus puissant de $H_1 = \{\theta \in \in \Theta_1\}$ contre $H_2 = \{\theta \in \Theta_2\}$ dans la classe K_ϵ.

Sous des conditions assez larges qui sont ordinairement remplies dans les problèmes, il existe une distribution la plus défavorable $\mathbf{Q}_1$ vérifiant les hypothèses du théorème 3. Il suffit d'exiger que Θ_1 soit compact et $f_\theta(x)$ continue par rapport à θ pour presque tous les x (pour plus de détails voir [50] et le chapitre V).

L'étude des liens entre les approches bayésienne et minimax sera poursuivie au § 9.

§ 6*. Tests sans biais

Dans ce paragraphe et dans le suivant, on se servira des principes d'absence de biais et d'invariance pour restreindre de façon naturelle la classe des tests envisagés, l'objectif de cette restriction étant la recherche des tests optimaux.

1. Définitions. Tests uniformément les plus puissants sans biais. Soit à éprouver comme dans le paragraphe précédent l'hypothèse multiple $H_1 = \{\theta \in \Theta_1\}$ contre $H_2 = \{\theta \in \Theta_2\}$ au vu d'un échantillon $X \in \mathbf{P}_\theta$, $\theta \in \Theta = \Theta_1 \cup \Theta_2$. Considérons un test π de classe $K_\epsilon = \{\pi : \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi \leqslant \epsilon\}$. Si par exemple Θ_1 est composé du seul point θ_1, $\mathbf{E}_{\theta_1}\pi = \epsilon$, alors ϵ est la probabilité de rejeter H_1 à tort. Une condition légitime que doit remplir le test π est que *la probabilité de rejeter H_1 à juste titre soit strictement supérieure à ϵ*. Le cas échéant il existerait des contre-hypothèses pour lesquels l'acceptation de H_1 est plus probable que dans les cas où H_1 est vraie. Une telle situation n'est pas souhaitable. Nous sommes conduits à la nécessité de distinguer l'importante classe de tests suivante.

DÉFINITION 1. On dit qu'un test π est *sans biais* ou *non biaisé* si

$$\inf_{\theta \in \Theta_2} \mathbf{E}_\theta \pi(X) \geqslant \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi(X). \tag{1}$$

Donc, un test $\pi \in K_\epsilon$ tel que $\sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi = \epsilon$ sera sans biais si $\beta_\pi(\theta) \geqslant \epsilon$ pour $\theta \in \Theta_2$. Désignons par $\check{K}_\epsilon$ la classe des tests sans biais de niveau $1 - \epsilon$.

Le test unilatéral π de région critique $T > c$ (ou $T < c$) pour les familles exponentielles, mentionné dans le paragraphe précédent, ne peut être un test sans biais de $H_1 = \{X \in \mathbf{P}_{\theta_1}\}$ contre $H_2 = \{X \in \mathbf{P}_{\theta_1}, \theta \neq \theta_1\}$, puisque $\Theta_2 = \{\theta : \theta \neq \theta_1\}$, $\mathbf{E}_\theta \pi < \epsilon$ pour $\theta < \theta_1$ si $\mathbf{E}_{\theta_1}\pi = \epsilon$ (cf. théorème 5.1).

En revânche, s'il existe des tests uniformément les plus puissants, ils

seront nécessairement sans biais, puisque leurs puissances $\beta(\theta)$, $\theta \in \Theta_2$, ne pourront être inférieures strictement à celle du test $\pi(X) \equiv \epsilon$.

Le principe d'absence de biais *) présente un intérêt en soi dans la mesure où il permet de restreindre naturellement la classe des tests. Ce qui nous donne la possibilité de construire des tests uniformément les plus puissants dans les classes $\check{K}_\epsilon$ dans les cas où ceux-ci n'existent pas dans la classe K_ϵ.

Nous verrons que ceci concerne en particulier le problème de test de l'hypothèse $H_1 = \{\theta \in [\theta_1, \theta_2]\}, \theta_1 \leqslant \theta_2$, contre l'alternative bilatérale $H_2 = \{\theta \notin [\theta_1, \theta_2]\}$ (comparer avec le n° 2 du § 5).

Pour chercher les tests uniformément les plus puissants sans biais on peut dans une large mesure se servir des méthodes déjà appliquées dont le contenu est exposé dans le lemme 5.2. Ceci étant, la proposition suivante peut nous être utile.

Supposons que des ensembles Θ_1 et Θ_2 de R^k admettent une frontière commune non vide Γ :

$$\Gamma = \partial\Theta_1 \cap \partial\Theta_2$$

($\partial\Theta_i$ désigne la frontière de Θ_i), c'est-à-dire l'ensemble des points adhérents simultanément à Θ_1 et à Θ_2. Supposons d'autre part que pour tous les $\pi \in \check{K}_\epsilon$

$$\beta_\pi(\theta) = \mathbf{E}_\theta \pi(X) = \epsilon \qquad \text{pour tous les } \theta \in \Gamma. \tag{2}$$

Cette propriété est visiblement toujours remplie si $\beta_\pi(\theta)$ dépend continûment de θ pour tout test π de $\check{K}_\epsilon$.

Comme

$$\beta_\pi(\theta) = \int \pi(x) f_\theta(x) \mu^n(dx), \qquad 0 \leqslant \pi(x) \leqslant 1,$$

cette fonction sera continue si $f_\theta(x)$ l'est par rapport à θ pour $[\mu^n]$-presque tous les x. Ceci résulte du corollaire 1 de l'Annexe VI.

Désignons par $\bar{K}_\epsilon$ la classe des tests π vérifiant (2).

LEMME 1. *Supposons que* $\check{K}_\epsilon \subset \bar{K}_\epsilon$ *(c'est-à-dire qu'est remplie* (2)*). Si* $\check{\pi}$ *est un test uniformément le plus puissant dans* $\bar{K}_\epsilon \cap K_\epsilon$, *il le sera dans* $\check{K}_\epsilon$.

DÉMONSTRATION. Il nous suffit de nous assurer que $\check{\pi} \in \check{K}_\epsilon$ et que $\check{K}_\epsilon \subset \bar{K}_\epsilon \cap K_\epsilon$. La deuxième de ces propositions résulte de l'hypothèse $\check{K}_\epsilon \subset$

*) Le terme « sans biais » a été utilisé aussi pour caractériser les estimateurs. La propriété d'absence de biais d'un estimateur est dans une certaine mesure identique à la propriété d'absence de biais d'un test : si un estimateur θ^* est à biais, alors $\mathbf{E}_{\theta_0}\theta^* \neq \theta_0$ et il existe d'autres valeurs du paramètre $\theta \neq \theta_0$ pour lesquelles $\mathbf{E}_\theta \theta^* = \theta_0$.

$\subset \bar{K}_\epsilon$. La première, du fait que le test $\pi \equiv \epsilon$ appartient à $\bar{K}_\epsilon \cap K_\epsilon$ et par suite $\inf_{\theta \in \Theta_2} \mathsf{E}_\theta \check{\pi}(X) \geqslant \inf_{\theta \in \Theta_2} \mathsf{E}_\theta \pi = \epsilon$. ◄

Le lemme 1 nous permet donc de ramener la recherche des tests uniformément les plus puissants sans biais à celle d'ordinaires tests uniformément les plus puissants mais sous les conditions aux limites (2). Si la frontière Γ est composée d'un nombre fini de points, on se retrouve dans les conditions du lemme 5.2 où il nous reste à vérifier que la fonction critique optimale obtenue π est indépendante de la valeur $\theta \in \Theta_2$ maximisant la fonctionnelle $\mathsf{E}_\theta \pi(X)$. Ceci exprimera que le test est uniformément le plus puissant.

Signalons maintenant le fait suivant lié à la dégénérescence des conditions (2), fait qu'il est plus simple d'illustrer en dimension un. Si $\Theta_1 = [\theta_1, \theta_2]$ et Θ_2 est le complémentaire de Θ_1, les conditions (2) auront la forme de deux équations $\mathsf{E}_{\theta_i} \pi(X) = \epsilon$, $i = 1, 2$. Ces équations se transforment en une seule dans le cas limite $\theta_1 = \theta_2$. Mais, le test π étant sans biais, sa puissance $\beta_\pi(\theta)$ doit prendre sa valeur minimale au point θ_1 (cf. (1)). Donc, si $\beta_\pi(\theta)$ est dérivable, le rôle des équations (2) pour $\theta_1 = \theta_2$ sera tenu par les égalités

$$\beta_\pi(\theta_1) = \epsilon, \quad \beta'_\pi(\theta_1) = 0. \tag{3}$$

Les conditions de dérivabilité de $\int f_\theta(x)\mu(dx)$, et partant de $\beta_\pi(\theta) = \mathsf{E}_\theta \pi(X)$ sont établies dans l'Annexe VI. Si ces conditions sont remplies, on a

$$\beta'_\pi(\theta) = \int \pi(x) f'_\theta(x) \mu^n(dx) = \int \pi(x) L'(x, \theta) f_\theta(x) \mu^n(dx) = \mathsf{E}_\theta \pi(X) L'(X, \theta).$$

Ceci exprime que les conditions (3) peuvent de nouveau être transcrites en termes d'intégrales :

$$\mathsf{E}_{\theta_1} \pi(X) = \epsilon, \quad \mathsf{E}_{\theta_1} \pi(X) L'(X, \theta_1) = 0. \tag{4}$$

Pour la famille exponentielle (5.9) par exemple, on a

$$L'(x, \theta) = c'(\theta)/c(\theta) + a'(\theta) T(x).$$

Comme $\mathsf{E}_\theta L'(X, \theta) = 0$, il vient $c'(\theta)/c(\theta) = -a'(\theta) \mathsf{E}_\theta T(X)$,

$$\mathsf{E}_\theta \pi(X) L'(X, \theta) = -a'(\theta) \mathsf{E}_\theta T(X) \cdot \mathsf{E}_\theta \pi(X) + a'(\theta) \mathsf{E}_\theta \pi(X) T(X),$$

et les équations (4) deviennent

$$\mathsf{E}_{\theta_1}(\pi(X) - \epsilon) = 0, \quad \mathsf{E}_{\theta_1}(\pi(X) - \epsilon) T(X) = 0.$$

A titre d'illustration considérons un cas pour l'examen duquel nous avons déjà préparé le terrain.

2. Alternatives bilatérales. Famille exponentielle.

THÉORÈME 1. *Supposons que $f_\theta(x)$ est définie par* (5.9) *et que l'on teste l'hypothèse $H_1 = \{\theta \in [\theta_1, \theta_2]\}$, $\theta_1 \leqslant \theta_2$, contre l'hypothèse $H_2 = \{\theta \notin [\theta_1, \theta_2]\}$. Si la fonction $a(\theta)$ est monotone, alors :*

1) *Dans la classe K_ϵ des tests sans biais de niveau* $1 - \epsilon$ *il existe un test uniformément le plus puissant $\check{\pi}$ de la forme*

$$\check{\pi}(x) = \begin{cases} 0 & si \quad c_1 < T(x) < c_2, \\ p_i & si \quad T(x) = c_i, i = 1, 2, \\ 1 & si \quad T(x) \notin [c_1, c_2], \end{cases} \tag{5}$$

où $T(x) = \sum_{i=1}^{n} U(x_i)$, et les constantes c_i, p_i, $i = 1, 2$, se déterminent à partir des conditions

$$\mathbf{E}_{\theta_i} \check{\pi}(X) = \epsilon, \quad i = 1, 2, \tag{6}$$

si $\theta_1 < \theta_2$ et des conditions

$$\mathbf{E}_{\theta_1} \check{\pi}(X) = \epsilon, \mathbf{E}_{\theta_1} (\check{\pi}(X) - \epsilon) T(X) = 0, \tag{7}$$

si $\theta_1 = \theta_2$.

2) *Le test $\check{\pi}$ minimise la fonction $\beta_\pi(\theta) = \mathbf{E}_\theta \pi(X)$ sous les conditions* (6) *à l'intérieur de l'intervalle $[\theta_1, \theta_2]$ et la maximise à l'extérieur de cet intervalle sous les conditions* (6) *ou* (7) (*pour $\theta_1 = \theta_2$ dans le dernier cas*).

3) *Pour $0 < \epsilon < 1$, $\theta_1 < \theta_2$, la fonction $\check{\beta}(\theta) = \mathbf{E}_\theta \check{\pi}(X)$ présente un minimum en un point $\theta_0 \in]\theta_1, \theta_2[$ et est strictement croissante lorsque θ s'éloigne de θ_0 vers la droite ou la gauche. Ceci étant, nous excluons le cas* (5.8).

Il est aisé de voir que ce théorème reprend pratiquement le théorème 5.2 à la seule différence que les assertions sont « inverses » et l'égalité $\theta_1 = \theta_2$ n'est pas exclue.

DÉMONSTRATION. Pour $\theta_1 < \theta_2$, la marche à suivre est exactement la même que pour le théorème 5.2. Dans la remarque 1 qui suivait ce théorème on a signalé que pour $\theta_1 < \theta_2$ tous les raisonnements restaient en vigueur si l'on teste l'hypothèse $\{\theta \notin [\theta_1, \theta_2]\}$ contre $\{\theta \in [\theta_1, \theta_2]\}$, c'est-à-dire, dans les notations de ce paragraphe, l'hypothèse H_2 contre H_1. Posons $\check{\pi}(x) = 1 - \pi^\circ(x)$, où π° est la fonction définie dans (5.6) sous les conditions $\mathbf{E}_{\theta_i} \pi^\circ(X) = 1 - \epsilon$, $i = 1, 2$, au lieu de (5.7). Il est alors évident que les propositions 2), 3) résultent directement des propositions correspondantes du théorème 5.2.

La première assertion du théorème découle de la seconde, puisque la classe des tests π vérifiant (6) est plus large que $\check{K}_\epsilon$, et par suite, $\check{\pi}$ maximi-

sera $\mathbf{E}_\theta \pi(X)$ dans la classe $\check{K}_\epsilon$ en tout point $\theta \notin [\theta_1, \theta_2]$. Ceci exprime que $\check{\pi}$ est un test uniformément le plus puissant sans biais.

Reste à traiter le cas $\theta_1 = \theta_2$. Il est plus simple d'appliquer le lemme 5.2. Prenons un point quelconque $\theta \neq \theta_1$ et cherchons le maximum de la quantité $\mathbf{E}_\theta \pi(X)$ sous les conditions

$$\mathbf{E}_{\theta_1} \pi(X) = \epsilon, \quad \mathbf{E}_{\theta_1} \pi(X) T(X) = \epsilon \mathbf{E}_{\theta_1} T(X). \tag{8}$$

On se retrouvera de toute évidence dans les conditions du lemme 5.2 si l'on pose $m = 2, f_1 = f_{\theta_1}, f_2 = Tf_{\theta_1}, f_3 = f_\theta, \epsilon_1 = \epsilon, \epsilon_2 = \epsilon \mathbf{E}_\theta T(X)$. En vertu de ce lemme, $\mathbf{E}_\theta \pi$ atteindra son maximum sur la fonction

$$\check{\pi}(x) = \begin{cases} 1 & \text{si} \quad f_\theta(x) > k_1 f_{\theta_1}(x) + k_2 T(x) f_{\theta_1}(x), \\ 0 & \text{si} \quad f_\theta(x) < k_1 f_{\theta_1}(x) + k_2 T(x) f_{\theta_1}(x). \end{cases}$$

Considérons la dernière inégalité qui peut être mise sous la forme

$$\frac{c(\theta)}{c(\theta_1)} e^{(a(\theta) - a(\theta_1))T(x)} < k_1 + k_2 T(x).$$

Il est clair que pour tous $c_1 < c_2$ on peut toujours choisir k_1 et k_2 de telle sorte que cette inégalité soit équivalente à

$$c_1 < T < c_2.$$

Ceci prouve que le test (5) maximise $\mathbf{E}_\theta \pi(X)$ sous les conditions (7) si seulement $c_i, p_i, i = 1, 2$, sont choisis dans (5) de façon à ce qu'on ait (7) (ou (8)). Ce test sera visiblement uniformément le plus puissant sans biais, puisque la classe des tests π vérifiant (8) est plus large que $\check{K}_\epsilon$, et par suite, $\check{\pi}$ maximisera $\mathbf{E}_\theta \pi(X)$ dans $\check{K}_\epsilon$ aussi. Pour parachever la démonstration du théorème, il nous reste donc à montrer que

LEMME 2. *Pour* $0 < \epsilon < 1$, *l'équation* (7) *admet une solution en* c_i *et* p_i, $i = 1, 2$.

DÉMONSTRATION. Nous prouverons ce lemme comme le lemme 5.1 en admettant pour simplifier que la $\mathbf{P}_{\theta_1}$-distribution de $T(X)$ est continue, c'est-à-dire que $\mathbf{P}_{\theta_1}(T(X) = c) = 0$ pour tous les c.

Rappelons que la densité de la distribution de T par rapport à une mesure ν peut être supposée égale à $g_\theta(t) = c(\theta)e^{\theta t}$ (cf. § 5). Les équations (7) et (8) seront alors équivalentes à

$$\mathbf{E}_{\theta_1}(1 - \pi(X)) = c(\theta_1) \int_{c_1}^{c_2} e^{\theta_1 t} \nu(dt) = 1 - \epsilon,$$

$$\mathbf{E}_{\theta_1}(1 - \pi(X))T(X) = c(\theta_1) \int_{c_1}^{c_2} t e^{\theta_1 t} \nu(dt) = (1 - \epsilon) c(\theta_1) \int_{c_1}^{c_2} t e^{\theta_1 t} \nu(dt). \tag{9}$$

En désignant $r(t) = t$, $m = \mathbf{E}_{\theta_1} T(X) = c(\theta_1) \int t e^{\theta_1 t} \nu(dt)$, on peut mettre les équations (9) sous la forme

$$c(\theta_1) \int_{c_1}^{c_2} e^{\theta_1 t} \nu(dt) = 1 - \epsilon,$$
$$c(\theta_1) \int_{c_1}^{c_2} r(t) e^{\theta_1 t} \nu(dt) = (1 - \epsilon) m. \tag{10}$$

Nous sommes arrivés à un problème confondu avec celui posé dans le lemme 5.1 à la seule différence que la distribution de densité $r(t) g_{\theta_1}(t)$ peut être une fonction généralisée (c'est-à-dire prendre des valeurs négatives). Dans ces nouvelles conditions, il faut poser $t_0 = m$. Pour le reste, les raisonnements du lemme 5.1 ne subissent pas de changement. ◄

§ 7*. Tests invariants

Dans ce paragraphe nous étudions une autre méthode de restriction de la classe des tests, basée cette fois-ci sur la notion d'invariance.

Supposons que $X \in \{\mathbf{P}_\theta\}$ et que $\{\mathbf{P}_\theta\}$ est une famille invariante. Rappelons les notations et les notions nécessaires (cf. § 2.19). Soit donné un groupe G de transformations mesurables g de $\mathscr{X}^n$ dans lui-même. On dit qu'une famille $\{\mathbf{P}_\theta\}$ est *invariante par le groupe G* si pour tout $g \in G$ et $\theta \in \Theta$ il existe un élément $\theta_g \in \Theta$ tel que

$$\mathbf{P}_{\theta_g}(X \in A) = \mathbf{P}_\theta(gX \in A),$$

quel que soit $A \in \mathfrak{B}^n_{\mathscr{X}}$.

Les transformations $\bar{g}$ de l'espace Θ définies par $\bar{g}\theta = \theta_g$ sous les conditions (A_0) forment un groupe $\bar{G}$ (cf. § 2.19).

DÉFINITION 1. On dira que le problème de test de l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ contre l'hypothèse $H_2 = \{\theta \in \Theta_2\}$, $\Theta_1 \cup \Theta_2 = \Theta$, est *invariant* si sont remplies les deux conditions suivantes :

1) La famille $\{\mathbf{P}_\theta\}$ est invariante par G.

2) Les ensembles Θ_1 et Θ_2 sont invariants par $\bar{g} \in \bar{G}$, c'est-à-dire que $\bar{g}\Theta_i = \Theta_i$, $i = 1, 2$.

Si un problème de test d'hypothèses est invariant, il est naturel de le résoudre à l'aide de tests invariants.

DÉFINITION 2. On dit qu'un test π est *invariant* si $\pi(X)$ est une statistique *) invariante par g :

$$\pi(gx) = \pi(x) \quad \text{pour tous les } x \in \mathscr{X}^n, \quad g \in G.$$

*) Voir note de la page 188.

Si π est un test non randomisé et Ω_j, la région d'acceptation de l'hypothèse H_j, l'invariance de π exprime que $g\Omega_j = \Omega_j, j = 1, 2$.

L'adéquation de l'usage des tests invariants se comprend le mieux sur des exemples. Au § 2.19 on trouvera une discussion générale sur l'interprétation de g comme un changement de variables et sur l'invariance des statistiques correspondantes par ce changement.

EXEMPLE 1. Les exemples les plus simples se rapportent au cas où le groupe $\bar{G}$ est trivial, c'est-à-dire que pour tout g la transformation $\bar{g}$ est la transformation identique $\bar{e}$ de l'espace Θ.

Supposons que $X \in \Phi_{0,\sigma^2}$ et soit à tester l'hypothèse $H_1 = \{\sigma \in [\sigma_1, \sigma_2]\}$ contre H_2. Dans ce cas

$$f_{\sigma^2}(X) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right\}.$$

Il est évident que la famille Φ_{0,σ^2} est invariante par le groupe G des transformations orthogonales g (les rotations) de l'espace $\mathscr{X}^n$, et de plus $\bar{g} = \bar{e}$ pour tout g. Il est donc naturel d'envisager des tests dépendant de la seule statistique $T(X) = \sum_{i=1}^{n} x_i^2$. Puisque $\sigma^{-2}T(X) \in \mathbf{\Gamma}_{1/2,n/2} = \mathbf{H}_n$, il vient $T(X) \in \mathbf{\Gamma}_{\alpha,n/2}$ pour $\alpha = 1/(2\sigma^2)$, et l'on est conduit à tester l'hypothèse $H_1 = \{\alpha \in [\alpha_1, \alpha_2]\}$, $\alpha_1 = 1/(2\sigma_2^2)$, $\alpha_2 = 1/(2\sigma_1^2)$, à l'aide de la statistique $T(X)$ dont la distribution $\mathbf{\Gamma}_{\alpha,n/2}$ appartient à une famille exponentielle. Les résultats des paragraphes précédents nous permettent de construire un test uniformément le plus puissant sans biais de niveau $1 - \epsilon$ qui est favorable à H_1 si

$$c_1 \leqslant T(X) \leqslant c_2, \tag{1}$$

où c_i sont choisis de telle sorte que $\mathbf{\Gamma}_{\alpha_1,n/2}(R \setminus [c_1, c_2]) = \mathbf{\Gamma}_{\alpha_2,n/2}(R \setminus [c_1, c_2]) = \epsilon$.

Signalons qu'on aurait pu construire le test (1) de cet exemple à l'aide du principe d'exhaustivité, puisque la statistique T est exhaustive. On sait en effet que toute l'information sur le paramètre σ^2 est concentrée dans T et le recours à d'autres statistiques (c'est-à-dire à une autre information sur l'échantillon) n'a pas de sens.

Dans la suite, partout où cela sera possible, on ramènera immédiatement le problème posé à un problème sur la distribution de statistiques exhaustives.

EXEMPLE 2. Soient $X \in \Phi_{\alpha,\sigma^2}$, $H_1 = \{\sigma \in [\sigma_1, \sigma_2]\}$. Alors $\theta = (\alpha, \sigma^2)$ et la translation $gX = X + c = (x_1 + c, \dots, x_n + c)$ induit la transformation

$\bar{g}\alpha = \alpha + c$ qui laisse invariante l'hypothèse H_1. Si on limite l'étude aux statistiques exhaustives

$$T_1 = \bar{x}, \quad T_2 = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

la transformation g nous donne

$$T_1(gX) = \bar{x} + c, \quad T_2(gX) = T_2(X).$$

La statistique T_2 est donc invariante par le groupe G. En d'autres termes, le test invariant π basé sur des statistiques exhaustives doit être une fonction de T_2. (On verra plus bas que *tout* test invariant π doit être fonction de T_2.) En vertu du § 2.32 on a $\sigma^{-2}T_2 \in \mathbf{\Gamma}_{1/2, (n-1)/2}$ et l'on est conduit au problème traité dans l'exemple précédent. Un test uniformément le plus puissant sans biais invariant sera de la forme $c_1 \leqslant T_2 \leqslant c_2$.

EXEMPLE 3. Les deux exemples envisagés plus haut faisaient intervenir une distribution normale. La distribution de l'échantillon X était une distribution normale multidimensionnelle de matrice des moments d'ordre deux diagonale. Il est utile de remarquer pour la suite que la famille des distributions normales multidimensionnelles Φ_{α, σ^2}, $\alpha \in R^m$, $\sigma^2 = \|\sigma_{ij}\|$, $i, j = 1, \ldots, m$, est invariante par le groupe G des transformations linéaires non dégénérées

$$gx = (x - a)C,$$

où C est une matrice inversible. En effet, nous devons nous assurer que pour une transformation g on a $\mathbf{P}_{\bar{g}\theta}(A) = \mathbf{P}_\theta(g^{-1}A)$, où $\mathbf{P}_\theta \equiv \Phi_{\alpha, \sigma^2}$, $\theta = (\alpha, \sigma^2)$, $g^{-1}A$ désigne comme toujours l'ensemble $\{x \in R^m : gx \in A\}$. On a ($\sigma = \sqrt{|\sigma^2|}$)

$$\Phi_{\alpha, \sigma^2}(g^{-1}A) = \frac{1}{(2\pi)^{m/2}\sigma} \int_{g^{-1}A} \exp\left\{-\frac{1}{2}(x - \alpha)\sigma^{-2}(x - \alpha)^T\right\} dx.$$

Le changement $y = gx$ nous donne

$$\Phi_{\alpha, \sigma^2}(g^{-1}A) = \\ = \frac{1}{(2\pi)^{m/2}\sigma|C|} \int_A \exp\left\{-\frac{1}{2}(g^{-1}y - \alpha)\sigma^{-2}(g^{-1}y - \alpha)^T\right\} dy.$$

Comme $g^{-1}y = yC^{-1} + a$, on peut mettre l'exposant de l'exponentielle de la dernière intégrale sous la forme

$$(y - (\alpha - a)C)C^{-1}\sigma^{-2}(C^{-1})^T(y - (\alpha - a)C)^T.$$

Si donc l'on pose

$$\bar{g}\theta = \bar{g}(\alpha, \sigma^2) = (g\alpha, C^T\sigma^2 C) = ((\alpha - a)C, C^T\sigma^2 C), \tag{2}$$

on obtient

$$\Phi_{\alpha, \sigma^2}(g^{-1}A) = \Phi_{\bar{g}(\alpha, \sigma^2)}(A). \tag{3}$$

EXEMPLE 4. Soient des hypothèses $H_j = \{X \in \mathbf{P}_{j, \alpha}\}$, $\alpha \in \mathscr{X}$, $j = 1, 2$, où $\mathbf{P}_{j, \alpha}$ sont des distributions de densités $f_j(x - \alpha)$, $j = 1, 2$. En d'autres termes, il nous faut déterminer le type de la distribution de X à une translation près. Il faut poser ici $\theta = (\nu, \alpha)$, $\nu = 1, 2$, $\alpha \in \mathscr{X}$, et considérer la transformation $gX = X + c$ qui induit la transformation $\bar{g}\theta = (\nu, \alpha + c)$ dans l'espace des paramètres. Il est clair que les hypothèses $H_j = \{\nu = j\}$, $j = 1, 2$, sont invariantes par $\bar{g}$, et par suite, le problème de test de ces hypothèses est invariant. La statistique

$$Y = (x_1 - x_n, \ldots, x_{n-1} - x_n)$$

est invariante par g (comparer avec le § 2.18). Sa distribution au point $y = (y_1, \ldots, y_{n-1})$ sous l'hypothèse H_j aura pour densité

$$f_j^Y(y) = \int \prod_{i=1}^{n-1} f_j(y_i + z) f_j(z) \mu(dz). \tag{4}$$

On voit que pour l'observation Y les hypothèses H_j se transforment en hypothèses simples en vertu desquelles les densités f_j^Y de Y sont de la forme (4). Dans ces conditions, on peut se servir du lemme de Neyman-Pearson et construire un test le plus puissant π qui nous fera accepter H_2 si

$$f_2^Y(Y)/f_1^Y(Y) > c. \tag{5}$$

Puisque ce test ne dépend pas de α, il sera un test uniformément le plus puissant de H_1 contre H_2 parmi les tests invariants basés sur la statistique Y.

D'après les exemples envisagés il est souhaitable d'être sûr que les autres tests invariants sont fonctions des statistiques invariantes choisies. Ceci concerne surtout le dernier exemple, puisque dans les deux précédents le choix des tests était guidé aussi par des considérations d'exhaustivité.

Introduisons quelques notions pour dégager les liens existant entre ces invariances. On dira que deux points x et x' de $\mathscr{X}^n$ sont *équivalents* par rapport à un groupe G s'il existe un $g \in G$ tel que $x' = gx$. Puisque G est un groupe, l'espace $\mathscr{X}^n$ est partitionné en classes d'équivalence disjointes appelées *orbites* dans le § 2.19. Pour obtenir une orbite, il suffit de prendre l'un quelconque de ses points x_0 et de lui appliquer toutes les transformations g de G. Pour les transformations orthogonales de l'exemple 1 les orbites sont des sphères centrées en l'origine des coordonnées.

Dire qu'une statistique T est invariante par G revient à dire qu'elle est constante sur chaque orbite.

DÉFINITION 3. On dit qu'une statistique T est un *invariant maximal* si elle est invariante et si de $T(x') = T(x)$ il s'ensuit que $x' = gx$ pour un certain $g \in G$.

Ceci exprime qu'un invariant maximal prend des valeurs différentes sur des orbites différentes.

THÉORÈME 1. *Soit T un invariant maximal. Une statistique S est invariante si et seulement si elle dépend de X par l'intermédiaire de T, c'est-à-dire s'il existe une fonction φ telle que $S(X) = \varphi(T(X))$.*

Pour simplifier l'exposé nous laisserons de côté l'importante question de la mesurabilité de φ. Signalons seulement que cette mesurabilité aura lieu dans les exemples envisagés dans ce paragraphe *).

DÉMONSTRATION. Si $S(x) = \varphi(T(x))$, on a $S(gx) = \varphi(T(gx)) = \varphi(T(x)) = S(x)$ et par suite S est invariante. Pour prouver la réciproque il faut montrer que $T(x) = T(x')$ entraîne $S(x) = S(x')$. En effet, la relation $T(x) = T(x')$ entraîne l'existence d'un g tel que $x' = gx$. Comme S est invariante, il vient $S(x) = S(x')$. ◄

Considérons à titre d'exemple le groupe G des translations

$$gx = x + c = (x_1 + c, \dots, x_n + c).$$

Nous avons déjà signalé que la statistique $Y(x) = (x_1 - x_n, \dots, x_{n-1} - x_n)$ était un invariant. Montrons que c'est un invariant maximal. En effet, de la relation $Y(x) = Y(x') \equiv (x_1' - x_n', \dots, x_{n-1}' - x_n')$ il s'ensuit que $x_i - x_n = x_i' - x_n'$ pour tous les $i = 1, \dots, n - 1$. En admettant que $x_n' - x_n = c$, on trouve que $x_i' = x_i + c$, $i = 1, \dots, n$, $x' = x + c = gx$, ce qui exprime que x' et x sont équivalents.

Nous pouvons retourner maintenant à l'exemple 3 et affirmer que le test (5) est uniformément le plus puissant de tous les tests invariants, puisque ces derniers sont, en vertu du théorème 1, des fonctions de Y et, par suite, il n'existe pas de test invariant plus puissant que (5).

En s'inspirant de ce qui précède, le lecteur peut s'assurer que la statistique $\sum_{i=1}^{n} x_i^2$ de l'exemple 1 est un invariant maximal.

S'il existe des statistiques exhaustives, il est plus commode de réduire le problème primitif d'abord à un problème sur la distribution de statistiques exhaustives et d'appliquer ensuite les considérations d'invariance comme

*) Pour plus de détails voir par exemple [50], [91].

dans l'exemple 2 où la statistique $T_2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$ est visiblement un invariant maximal dans l'observation $(\bar{x}, T_2)$.

Signalons une fois de plus, en conclusion de ce paragraphe, que l'approche liée à l'invariance consiste à réduire les problèmes de test d'hypothèses à des problèmes plus simples relatifs à la distribution d'invariants maximaux. Dans les nouvelles conditions qui sont plus simples, il est souvent possible de construire un test le plus puissant ou un test uniformément le plus puissant. De ce point de vue le « principe d'invariance » est voisin des « principes » d'exhaustivité et d'absence de biais en vertu desquels le problème primitif se réduit à un problème portant sur des statistiques exhaustives ou sans biais.

§ 8*. Lien avec les régions de confiance

1. Lien entre les tests et les régions de confiance. Lien entre les propriétés d'optimalité. Les notions de région de confiance et de test sont étroitement liées entre elles. Rappelons la définition de la région de confiance qui a été donnée au § 2.31.

Soit $X \in \mathbf{P}_\theta, \theta \in \Theta$.

DÉFINITION 1. On dit qu'un sous-ensemble aléatoire $\Theta^* = \Theta^*(X, \epsilon)$ d'un espace de paramètres Θ est une *région de confiance au seuil* $1 - \epsilon$ si

$$\mathbf{P}_\theta(\Theta^*(X, \epsilon) \ni \theta) \geqslant 1 - \epsilon \tag{1}$$

pour tous les $\theta \in \Theta$.

Il est évident que la région de confiance généralise l'intervalle de confiance. La signification est la même : la région de confiance recouvre la véritable valeur du paramètre avec une probabilité $\geqslant 1 - \epsilon$.

Désignons

$$\Omega(\theta, \epsilon) = \{x \in \mathscr{X}^n : \quad \theta \in \Theta^*(x, \epsilon)\}. \tag{2}$$

Les relations

$$\theta \in \Theta^*(x, \epsilon) \quad \text{et} \quad x \in \Omega(\theta, \epsilon) \tag{3}$$

seront alors équivalentes.

La définition de la région de confiance suppose que l'ensemble $\Omega(\theta, \epsilon)$ de (2) est mesurable, si bien que la probabilité de (1) a un sens et est égale à $\mathbf{P}_\theta(X \in \Omega(\theta, \epsilon))$.

Les régions de confiance et les tests de l'hypothèse $H_1 = \{\theta = \theta_1\}$ contre l'hypothèse concurrente $H_2 = \{\theta \in \Theta_2\}$, $\theta_1 \notin \Theta_2$, sont reliés entre eux de la manière suivante. Supposons que pour chaque θ_1 est défini l'ensemble $\Theta_2 = \Theta_2(\theta_1) \not\ni \theta_1$.

THÉORÈME 1. 1) *Considérons pour chaque* θ_1 *un test non randomisé* $\pi = \delta$ *de niveau* $1 - \epsilon$ *de l'hypothèse* H_1 *contre* H_2 *et désignons par* $\Omega(\theta_1, \epsilon)$ *la région d'acceptation de* H_1. *Alors l'ensemble*

$$\Theta^*(X, \epsilon) = \{\theta \in \Theta : X \in \Omega(\theta, \epsilon)\}$$

sera une région de confiance au seuil $1 - \epsilon$.

Réciproquement, si $\Theta^*(X, \epsilon)$ *est une région de confiance au seuil* $1 - \epsilon$, *l'ensemble* $\Omega(\theta_1, \epsilon) \subset \mathscr{X}^n$ *défini dans* (2) *et pris pour région d'acceptation de* H_1 *déterminera un test de* $H_1 = \{\theta = \theta_1\}$ *contre* $H_2 = \{\theta \in \Theta_2(\theta_1)\}$ *de niveau* $1 - \epsilon$ *pour tout* $\Theta_2(\theta_1)$, $\theta_1 \notin \Theta_2(\theta_1)$.

2) *Si un test* π *de région d'acceptation* $\Omega(\theta_1, \epsilon)$ *pour* H_1 *est uniformément le plus puissant, l'ensemble correspondant* $\Theta^*(X, \epsilon)$ *minimise la probabilité*

$$\mathbf{P}_\theta(\theta' \in \Theta^*(X, \epsilon)) \quad \textit{pour tous les } \theta, \ \theta', \theta \in \Theta_2(\theta'), \tag{4}$$

dans la classe des régions de confiance au seuil $1 - \epsilon$.

La proposition réciproque est vraie : la minimalité de (4) *exprime que l'ensemble correspondant* $\Omega(\theta, \epsilon)$ *définit un test uniformément le plus puissant.*

Si le paramètre θ est scalaire, les cas les plus fréquents sont : $\Theta_2(\theta') = \{\theta : \theta \neq \theta'\}$ et $\Theta_2(\theta') = \{\theta : \theta > \theta'\}$ (ou $\{\theta : \theta < \theta'\}$). Dans (4) on aura affaire à une minimisation pour tous les $\theta' \neq \theta$ pour le premier cas, et pour tous les $\theta' < \theta$ pour le second.

Ainsi dans (4) le théorème affirme qu'est minimisée la probabilité $\mathbf{P}_\theta$ que dans la région de confiance Θ^* tombe n'importe quelle autre valeur $\theta' \neq \theta$ telle que $\theta \in \Theta_2(\theta')$. Ceci nous fournit un procédé de mise en évidence des intervalles de confiance optimaux.

DÉFINITION 2. Les régions de confiance pour lesquelles est minimisée (4) sous la condition (1) s'appellent *régions de confiance les plus exactes* (au seuil $1 - \epsilon$) *pour les alternatives* θ' *telles que* $\theta \in \Theta_2(\theta')$.

Une justification supplémentaire de cette notion d'optimalité de l'intervalle de confiance sera donnée plus bas.

Le théorème 1 exprime donc que l'« inversion » de l'ensemble $\Omega(\theta_1, \epsilon)$ pour les tests uniformément les plus puissants fournit la région de confiance la plus exacte. Ceci étant, il est important de remarquer que cette procédure de construction des régions de confiance n'est pas liée à la dimension de θ. On peut également envisager des paramètres θ de dimension infinie et les identifier avec la distribution $\mathbf{P}$ de X. Les relations d'équivalence (3), où $\Omega(\theta, \epsilon) = \Omega(\mathbf{P}, \epsilon)$ est la région d'acceptation de l'hypothèse $\{X \in \mathbf{P}\}$, l'alternative étant $\{X \in \mathbf{P}_1 \neq \mathbf{P}\}$, nous permettent de construire la région de confiance pour $\mathbf{P}$. Par exemple, nous avons vu au § 1.6 que la

statistique $D_n = \sqrt{n} \sup_t |F_n^*(t) - F(t)|$ sous la condition $X \in \mathbf{P}$, où F est fonction de répartition continue de $\mathbf{P}$, ne dépend pas de F et peut être déterminée. Nous pouvons donc trouver un $d = d(\epsilon)$ tel que $\mathbf{P}(D_n \leqslant d(\epsilon)) = 1 - \epsilon$. Donc, la relation

$$\sqrt{n} \sup_t |F_n^*(t) - F(t)| \leqslant d$$

définit la région d'acceptation de l'hypothèse $\{X \in \mathbf{P}\}$ pour un test de niveau $1 - \epsilon$.

Mais cette relation définit aussi la région de confiance pour F pour la simple raison qu'aucune procédure spéciale d'« inversion » n'est nécessaire, puisque cette inégalité est symétrique par rapport à F et F_n^*.

DÉMONSTRATION du théorème 1. Elle coule presque de source et s'appuie sur l'équivalence (3) en vertu de laquelle

$$\mathbf{P}_\theta(\theta \in \Theta^*(X, \epsilon)) = \mathbf{P}_\theta(X \in \Omega(\theta, \epsilon)) \geqslant 1 - \epsilon.$$

Ceci prouve la première proposition. Pour établir la seconde, considérons une autre région de confiance $\tilde{\Theta}^*(X, \epsilon)$. Soit $\tilde{\Omega}(\theta, \epsilon)$ le sous-ensemble correspondant de $\mathscr{X}^n$. On a alors

$$\mathbf{P}_\theta(X \in \tilde{\Omega}(\theta, \epsilon)) = \mathbf{P}_\theta(\theta \in \tilde{\Theta}^*(X, \epsilon)) \geqslant 1 - \epsilon,$$
$$\mathbf{P}_\theta(X \in \tilde{\Omega}(\theta_1, \epsilon)) \geqslant \mathbf{P}_\theta(X \in \Omega(\theta_1, \epsilon))$$

pour $\theta \in \Theta_2(\theta_1)$ et par suite

$$\mathbf{P}_\theta(\theta_1 \in \tilde{\Theta}^*(X, \epsilon)) \geqslant \mathbf{P}_\theta(\theta_1 \in \Theta^*(X, \epsilon)). \blacktriangleleft$$

Considérons maintenant un cas particulier faisant intervenir un paramètre scalaire θ.

2. Intervalles de confiance les plus exacts.

THÉORÈME 2. *Supposons que l'ensemble $\Omega(\theta, \epsilon)$ du test uniformément le plus puissant étudié dans le théorème 1 est de la forme*

$$c_1(\theta, \epsilon) \leqslant T(x) \leqslant c_2(\theta, \epsilon),$$

*où $c_i(\theta, \epsilon)$ dépendent de façon monotone et continue *) de θ. Supposons de plus, pour fixer les idées, que $c_i(\theta, \epsilon)$ sont strictement croissantes. Alors la région de confiance la plus exacte (au seuil $1 - \epsilon$) pour les contre-hypothèses θ' telles que $\theta \in \Theta_2(\theta')$ sera un intervalle de la forme*

$$c_2^{-1}(T, \epsilon) \leqslant \theta \leqslant c_1^{-1}(T, \epsilon),$$

---

*) Les propriétés de monotonie et de continuité de $c_i(\theta, \epsilon)$ résultent généralement des mêmes propriétés de la fonction de répartition $\mathbf{P}_\theta(T(X) < c)$. Dans les notations du § 2.31, $c_1(\theta, \epsilon) = G_\theta^{-1}(\epsilon_1)$, $c_2(\theta, \epsilon) = G_\theta^{-1}(1 - \epsilon_2)$, où G_θ est la fonction de répartition de $T(X)$, $\epsilon_1 + \epsilon_2 = \epsilon$.

où $T = T(X)$ *et* $c_i^{-1}(t, \epsilon)$ *sont les solutions des équations* $c_i(\theta, \epsilon) = t$ *par rapport à* θ.

Nous voyons donc que la procédure de construction de l'intervalle de confiance est au fond la même que dans le § 2.31 à la seule différence que la statistique S est remplacée ici par la statistique T d'un test uniformément le plus puissant.

La démonstration du théorème qui est évidente est laissée au soin du lecteur.

Considérons maintenant plus en détail les *intervalles de confiance unilatéraux pour un* θ *scalaire*. On se sert de ces intervalles pour estimer le paramètre unilatéralement. Ces situations se présentent lorsqu'on estime la probabilité d'un événement indésirable ou, par exemple, la valeur de l'effort de rupture d'un nouvel alliage.

Pour raison de symétrie on peut se limiter à l'étude de la borne de confiance inférieure $\theta^-(X, \epsilon)$ pour laquelle

$$\mathbf{P}_\theta(\theta^-(X, \epsilon) \leqslant \theta \geqslant 1 - \epsilon. \tag{5}$$

DÉFINITION 3. On appelle *borne de confiance inférieure la plus exacte au seuil* $1 - \epsilon$ la borne $\theta^- = \theta^-(X, \epsilon)$ telle que $\mathbf{P}_\theta(\theta^- \leqslant \theta')$ soit minimale pour tout $\theta' < \theta$.

Supposons que $w(\theta^-, \theta)$ est une mesure des pertes entraînées par une « sous-estimation » de θ : $w(\theta^-, \theta) = 0$ pour $\theta^- \geqslant \theta$, et $w(\theta^-, \theta) \geqslant 0$ pour $\theta^- < \theta$; ceci étant, $w(\theta^-, \theta)$ croît continûment lorsque θ^- s'éloigne de θ, et $\mathbf{E}_\theta w(\theta^-, \theta) < \infty$.

La proposition suivante éclaire dans une certaine mesure la définition 3.

LEMME 1. *La borne inférieure la plus exacte* θ^- *minimise* $\mathbf{E}_\theta w(\theta^-, \theta)$ *sous la condition* (5) *et pour toute fonction w possédant les propriétés mentionnées ci-dessus.*

DÉMONSTRATION. Soit $\tilde{\theta}^-$ une autre borne inférieure. Les accroissements de $d_u w(u, \theta)$ par rapport à u étant strictement négatifs dans le domaine $u < \theta$, il vient

$$\mathbf{E}_\theta w(\theta^-, \theta) = \int_{-\infty}^{\theta} w(u, \theta) d_u \mathbf{P}_\theta(\theta^- < u) = -\int_{-\infty}^{\theta} \mathbf{P}_\theta(\theta^- < u) d_u w(u, \theta) \leqslant$$

$$\leqslant -\int_{-\infty}^{\theta} \mathbf{P}_\theta(\tilde{\theta}^- < u) d_u w(u, \theta) = \mathbf{E}_\theta w(\tilde{\theta}^-, \theta). \blacktriangleleft$$

Nous voyons donc que la manière dont les régions de confiance les plus exactes ont été définies dans le cas d'intervalles unilatéraux est tout à fait

naturelle. On peut maintenant utiliser les théorèmes 1, 2 et les résultats du § 5 pour construire des intervalles de confiance unilatéraux sous forme explicite dans le cas où le rapport de vraisemblance est monotone.

THÉORÈME 3. *Soit* $X \in \mathbf{P}_\theta$ *et supposons que la famille* $\{\mathbf{P}_\theta\}$ *possède un rapport de vraisemblance monotone pour une statistique* $T(X)$ *dont la* $\mathbf{P}_\theta$*-distribution* $G_\theta(t) = \mathbf{P}_\theta(T(X) < t)$ *est continue par rapport à* θ *et* t. *Alors la statistique* T *dépend de façon monotone et continue de* θ *(c'est-à-dire que* $G_\theta(t)$ *décroît continûment lorsque* θ *croît*, cf. définition 2.31.3). *Si* $b(t, \gamma)$ *est la solution de l'équation* $G_\theta(t) = \gamma$ *par rapport à* θ, *la borne inférieure la plus exacte* $\theta^-(X, \epsilon)$ *au seuil* $1 - \epsilon$ *est égale à*

$$\theta^-(X, \epsilon) = b(T(X), 1 - \epsilon).$$

En d'autres termes, dans le théorème 2.31.1 on obtient la borne de confiance inférieure la plus exacte si pour S on prend la statistique T.

DÉMONSTRATION. Dans notre cas il faut poser $\Theta_2(\theta) = \{t : t > \theta\}$ dans les hypothèses des théorèmes 1 et 2. Le théorème 5.1 affirme l'existence d'un test uniformément le plus puissant non randomisé de $H_1 = \{\theta = \theta_1\}$ contre $H_2 = \{\theta > \theta_1\}$, de région d'acceptation $\Omega(\theta_1, \epsilon) = \{X : T(X) < c\}$ de H_1, où $c = c(\theta_1, 1 - \epsilon) = G_{\theta_1}^{-1}(1 - \epsilon)$ se détermine à partir de la condition

$$\mathbf{P}_{\theta_1}(T(X) < c(\theta_1, 1 - \epsilon)) = 1 - \epsilon.$$

Ceci étant,

$$\mathbf{P}_\theta(T(X) \geqslant c) > \epsilon = \mathbf{P}_{\theta_1}(T(X) \geqslant c)$$

pour $\theta > \theta_1$. La dernière relation exprime que $c(\theta_1, 1 - \epsilon) < c(\theta, 1 - \epsilon)$ pour $\theta_1 < \theta$, c'est-à-dire que la fonction $c(\theta, 1 - \epsilon)$ est strictement croissante par rapport à θ. La continuité de $c(\theta, 1 - \epsilon) = G_\theta^{-1}(1 - \epsilon)$ par rapport à θ résulte de celle de G_θ.

Nous voyons que les conditions des théorèmes 1 et 2 sont remplies pour $c_2(\theta, \epsilon) = c(\theta, 1 - \epsilon)$, et par suite, la région de confiance la plus exacte est l'intervalle $]c^{-1}(T(X), 1 - \epsilon), \infty[$, où, comme déjà vu au théorème 2.31.1, $c^{-1}(T, 1 - \epsilon) = b(T, 1 - \epsilon)$. ◄

On construirait exactement de la même façon la borne supérieure la plus exacte $\theta^+(X, \epsilon)$.

Supposons maintenant que $\theta^-(X, \epsilon_1) < \theta^+(X, \epsilon_2)$ sont des bornes de confiance inférieure et supérieure au seuil $1 - \epsilon_1$ et $1 - \epsilon_2$ respectivement. Puisque les événements $\{\theta^-(X, \epsilon_1) > \theta\}$ et $\{\theta^+(X, \epsilon_2) < \theta\}$ sont disjoints, on a

$$\mathbf{P}_\theta(\theta^-(X, \epsilon_1) < \theta < \theta^+(X, \epsilon_2)) = 1 - \epsilon_1 - \epsilon_2,$$

et $]\theta^-(X, \epsilon_1), \theta^+(X, \epsilon_2)[$ est un intervalle de confiance au seuil $1 - \epsilon_1 - \epsilon_2$.

Soient $w_1(\theta^-, \theta)$ et $w_2(\theta^+, \theta)$ des fonctions de perte pour les bornes θ^- et θ^+, possédant les propriétés décrites dans le lemme 1.

LEMME 2. *Supposons que $w(\theta^-, \theta^+, \theta) = w_1(\theta^-, \theta) + w(\theta^+, \theta)$. Alors l'intervalle de confiance $]\theta^-, \theta^+[$ formé par les bornes supérieure et inférieure les plus exactes minimise $\mathbf{E}_\theta w(\theta^-, \theta^+, \theta)$ sous les conditions*

$$\mathbf{P}_\theta(\theta^- > \theta) \leqslant \epsilon_1, \quad \mathbf{P}_\theta(\theta^+ < \theta) \leqslant \epsilon_2.$$

Ce lemme est une conséquence évidente du lemme 1. Il indique que l'intervalle de confiance construit à l'aide des bornes inférieure et supérieure les plus exactes sera aussi optimal.

Le théorème 3 nous permet de construire de tels intervalles sous une forme explicite pour les familles paramétriques de distributions à rapport de vraisemblance monotone.

Nous proposons au lecteur de s'assurer à l'aide des remarques faites que les intervalles de confiance construits au § 2.32 pour la moyenne et la variance de la distribution normale admettront des bornes inférieures et supérieures les plus exactes.

Dans le théorème 1 et dans les considérations ultérieures on a supposé que le test uniformément le plus puissant n'était pas randomisé. Cette condition n'est cependant pas essentielle. Tout test randomisé π peut être représenté comme un test non randomisé par l'introduction d'une observation supplémentaire Y indépendante de X et uniformément distribuée sur $[0, 1]$. En effet, considérons pour le nouvel échantillon (X, Y) la région critique

$$\Omega = \{(x, y) : \pi(x) \geqslant y\},$$

c'est-à-dire posons $\delta(X, Y) = 1$ si $(X, Y) \in \Omega$, et $\delta(X, Y) = 0$ sinon. Pour toute distribution de X on a alors

$$\mathbf{P}(\delta(X, Y) = 1) = \mathbf{P}(\pi(X) \geqslant Y) = \int_0^1 \mathbf{P}(\pi(X) \geqslant y)dy = \mathbf{E}\pi(X),$$

et par suite le test δ est équivalent à π par ses paramètres.

Comment utiliser cette circonstance pour construire des intervalles de confiance dans les hypothèses du théorème 3? Supposons pour simplifier que la statistique $T(X)$ est à valeurs entières (nous avons vu que l'absence de tests uniformément les plus puissants non randomisés est due au seul fait que la distribution de T est discrète). L'observation $S(X, Y) = T(X) + Y$, $Y \in \mathbf{U}_{0,1}$ conserve alors toute l'information contenue dans $T(X)$, puisque $T(X)$ est la partie entière de $S(X, Y)$. En choisissant un $c(\theta, \epsilon)$ non entier, la recette du test uniformément le plus puissant de niveau $1 - \epsilon$ sera la suivante : accepter l'hypothèse H_1 si

$$S(X, Y) \leqslant c(\theta_1, 1 - \epsilon).$$

Nous avons ainsi construit les ensembles $\Omega(\theta, \epsilon)$. Il ne reste maintenant qu'à les « inverser » à l'aide du même procédé qu'avant. Nous obtenons la borne inférieure

$$\theta^-(X, Y, \epsilon) = c^{-1}(T(X) + Y, 1 - \epsilon),$$

où c^{-1} est la fonction inverse de c par rapport au premier argument. A la forme d'écriture même, on voit que pour déterminer θ^- il faut effectuer une observation supplémentaire Y.

EXEMPLE 1. Soit $X \in \mathbf{B}_p$. Supposons qu'on s'intéresse à la borne supérieure p^+ au seuil $1 - \epsilon$ pour la probabilité $p = \mathbf{P}(x_i = 1) = 1 - \mathbf{P}(x_i = 0)$. La famille de distributions $\{\mathbf{B}_p\}$ est exponentielle et vérifie les conditions du théorème 3, où il faut poser $T(X) = \sum_{i=1}^{n} x_i$. Considérons l'observation

$$S = \sum_{i=1}^{n} x_i + Y, \qquad Y \in \mathbf{U}_{0,1}.$$

Sa densité en un point $t \in [0, n + 1]$ est $C_n^{[t]} p^{[t]} (1 - p)^{n-[t]}$. Notons $G_p(t)$ la fonction de répartition de cette densité. Dans ces conditions, p^+ sera solution de l'équation $G_p(t) = \epsilon$.

3. Régions de confiance sans biais. Revenons maintenant aux régions de confiance les plus exactes. Le théorème 3 nous permet de déterminer les bornes inférieures et supérieures les plus exactes en nous basant sur le fait que dans bien des cas il existe un test uniformément le plus puissant pour les hypothèses unilatérales $\{\theta > \theta_1\}$ et $\{\theta < \theta_1\}$ concurrentes de l'hypothèse de base $\{\theta = \theta_1\}$. Si l'on essaye d'appliquer directement les théorèmes 1 et 2 à la construction des intervalles de confiance les plus exacts, il faudra exiger l'existence d'un test uniformément le plus puissant de l'hypothèse $\{\theta = \theta_1\}$ contre l'hypothèse $\{\theta \neq \theta_1\}$, ce qui est très rare. L'issue est de restreindre naturellement la classe des intervalles de confiance envisagés d'après le même principe que pour les classes des tests étudiés dans les §§ 6, 7. Plus exactement, introduisons les notions de régions de confiance invariantes et sans biais.

Supposons comme précédemment qu'à tout θ est associé un ensemble $\Theta_2(\theta)$, $\theta \notin \Theta_2(\theta)$.

DÉFINITION 4. On dit qu'une région de confiance $\Theta^*(X, \epsilon)$ pour θ au seuil $1 - \epsilon$ est *sans biais pour les contre-hypothèses θ' telles que* $\theta \in \Theta_2(\theta')$ si

$$\mathbf{P}_\theta(\theta' \in \Theta^*(X, \epsilon)) \leqslant 1 - \epsilon \quad \text{pour tous les } \theta, \theta', \theta \in \Theta_2(\theta'). \tag{6}$$

La région $\Theta^*(X, \epsilon)$ est dite simplement *sans biais* si (6) est vraie pour tous les $\theta' \neq \theta$.

L'absence de biais pour la région de confiance exprime que *la probabi-*

lité de recouvrir une fausse valeur de θ' est au plus égale à celle de recouvrir la vraie valeur.

DÉFINITION 5. Les régions de confiance pour lesquelles (4) est minimisée sous les conditions (1), (6) s'appellent *régions de confiance sans biais les plus exactes* (au seuil $1 - \epsilon$) pour les contre-hypothèses telles que $\theta \in \in \Theta_2(\theta')$.

THÉORÈME 4. 1) *Les tests non randomisés sans biais engendrent, en vertu de l'équivalence* (3), *des régions de confiance sans biais, et réciproquement.*

2) *Si $\Omega(\theta_1, \epsilon)$ est pour chaque $\theta_1 \in \Theta$ la région d'acceptation de l'hypothèse $\{\theta = \theta_1\}$ contre l'hypothèse $\{\theta \in \Theta_2(\theta_1)\}$ par un test uniformément le plus puissant non randomisé sans biais, l'ensemble correspondant $\Theta^*(X, \epsilon)$ sera la région de confiance sans biais la plus exacte, et réciproquement.*

DÉMONSTRATION. Elle répète intégralement celle du théorème 1 à laquelle il faut ajouter seulement que la propriété d'absence de biais se conserve lorsqu'on passe des tests aux régions de confiance et réciproquement. En effet, les relations (1) et (6) sont équivalentes à

$$\sup_{\theta\in\Theta_2(\theta_1)} \mathbf{P}_\theta(X \in \Omega(\theta_1, \epsilon)) \leqslant 1 - \epsilon \leqslant \mathbf{P}_{\theta_1}(X \in \Omega(\theta_1, \epsilon)).$$

Si $\pi(X)$ est la fonction critique des tests non randomisés du théorème ($\pi(X) = 0$ si $X \in \Omega(\theta_1, \epsilon)$), on obtient

$$\mathbf{E}_\theta \pi(X) = 1 - \mathbf{P}_\theta(X \in \Omega(\theta_1, \epsilon)),$$
$$\inf_{\theta\in\Theta_2(\theta_1)} \mathbf{E}_\theta \pi(X) \geqslant \epsilon \geqslant \mathbf{E}_{\theta_1} \pi(X).$$

Ce qui est visiblement la propriété d'absence de biais équivalente à (6). ◄

Si l'on se sert des résultats du § 6 pour construire la région de confiance sans biais la plus exacte pour le paramètre θ d'une famille exponentielle, on obtiendra le même intervalle $]\theta^-, \theta^+[$ qu'avec un rapport de vraisemblance monotone, c'est-à-dire un intervalle dont θ^- et θ^+ sont respectivement les bornes inférieure et supérieure les plus exactes au seuil $1 - \epsilon/2$.

4. Régions de confiance invariantes. La définition suivante utilise les notations et notions du paragraphe précédent. Soit $\{\mathbf{P}_\theta\}$ une famille de distributions invariante par G.

DÉFINITION 6. On dit qu'une région de confiance $\Theta^*(X, \epsilon)$ est *invariante* *) *par un groupe* G si

$$\Theta^*(gX, \epsilon) = \bar{g}\Theta^*(X, \epsilon) \tag{7}$$

pour tous les $g \in G$.

*) Si l'on s'en tient à la remarque faite dans la note de la page 188, il serait plus logique d'appeler *équivariante* une région de confiance vérifiant (7).

Cette notion a la même signification que celle d'un estimateur équivariant (§ 2.19). Si les applications g et $\bar{g}$ sont traitées comme un changement de système de coordonnées préserveant la distribution, alors (7) exprime que la région de confiance ne dépend pas du système de coordonnées dans lequel sont exprimées les données initiales.

DÉFINITION 7. On dit que $\Theta^*(X, \epsilon)$ est une *région de confiance invariante la plus exacte* au seuil $1 - \epsilon$ si est minimisée $\mathbf{P}_\theta(\theta' \in \Theta^*(X, \epsilon))$ pour tous les $\theta' \neq \theta$ dans la classe de tous les Θ^* vérifiant (7) et la condition $\mathbf{P}_\theta(\theta \in \Theta^*(X, \epsilon)) = 1 - \epsilon$.

Supposons que $\Omega(\theta_1, \epsilon)$ est la région d'acceptation de l'hypothèse $H_1 = \{\theta = \theta_1\}$ contre l'hypothèse $\{\theta \neq \theta_1\}$ pour un test invariant de niveau $1 - \epsilon$. Signalons qu'il existe une différence fondamendale entre les définitions d'un test invariant et d'une région de confiance invariante (cette différence n'existerait pas s'il fallait que $g\Omega(\theta, \epsilon) = \Omega(\bar{g}\theta, \epsilon)$ et non pas $g\Omega(\theta, \epsilon) = \Omega(\theta, \epsilon)$). De ce fait la correspondance entre les tests invariants uniformément les plus puissants et les intervalles de confiance invariants les plus exacts est plus compliquée que dans les théorèmes précédents.

Considérons un groupe d'applications G et supposons que pour tout θ_1 il existe un sous-groupe $G[\theta_1]$ de G laissant invariant le problème de test de l'hypothèse $H_1 = \{\theta = \theta_1\}$. En d'autres termes, $\bar{g}\theta_1 = \theta_1$ pour $g \in G[\theta_1]$.

THÉORÈME 5. *Supposons que $\Theta^*(X, \epsilon)$ est une région de confiance au seuil $1 - \epsilon$ invariante par G. Alors :*

1) *La région $\Omega(\theta, \epsilon) = \{x : \theta \in \Theta^*(x, \epsilon)\}$ sera invariante par $G[\theta]$ pour chaque θ.*

2) *Si la région $\Omega(\theta_1, \epsilon)$ qui correspond à $\Theta^*(X, \epsilon)$ est la région d'acceptation de H_1 contre l'hypothèse $\{\theta \neq \theta_1\}$ pour un test uniformément le plus puissant, invariant, de niveau $1 - \epsilon$, alors $\Theta^*(X, \epsilon)$ sera la région de confiance invariante la plus exacte.*

DÉMONSTRATION. 1) Soit $g \in G[\theta]$. Alors $\bar{g}\theta = \theta$,

$$g\Omega(\theta, \epsilon) = \{gx : \theta \in \Theta^*(x, \epsilon)\} = \{x : \theta \in \Theta^*(g^{-1}x, \epsilon)\} =$$
$$= \{x : \theta \in \bar{g}^{-1}\Theta^*(x, \epsilon)\} = \{x : \bar{g}\theta \in \Theta^*(x, \epsilon)\} =$$
$$= \{x : \theta \in \Theta^*(x, \epsilon)\} = \Omega(\theta, \epsilon).$$

2) Soit $\tilde{\Theta}^*$ une autre région de confiance invariante au seuil $1 - \epsilon$. En vertu de la première proposition, il lui est associé un test invariant de niveau $1 - \epsilon$ de région d'acceptation $\tilde{\Omega}(\theta_1, \epsilon)$ de H_1.

Puisque par hypothèse

$$\mathbf{P}_\theta(X \in \Omega(\theta_1, \epsilon)) \geqslant \mathbf{P}_\theta(X \in \tilde{\Omega}(\theta_1, \epsilon)),$$

il vient

$$\mathbf{P}_\theta(\theta_1 \in \Theta^*(X, \epsilon)) \geqslant \mathbf{P}_\theta(\theta_1 \in \tilde{\Theta}^*(X, \epsilon))$$

pour $\theta_1 \neq \theta$. ◄

EXEMPLE 2. Soit $X \in \Phi_{\alpha, \sigma^2}$. On demande de déterminer la région de confiance la plus exacte pour le paramètre σ^2, α étant inconnu. Nous avons vu dans l'exemple 2 du paragraphe précédent que la famille Φ_{α, σ^2} était invariante par les translations $gX = X + c$ si $\bar{g}(\alpha, \sigma^2) = (\alpha + c, \sigma^2)$. La statistique $S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\mathrm{x}_i - \bar{\mathrm{x}})^2$ est un invariant maximal construit à l'aide d'une statistique exhaustive. De plus, l'hypothèse $H_1 = \{\sigma = \sigma_1\}$ est invariante par G. Conformément à l'exemple 7.2, un test uniformément le plus puissant, invariant, sans biais pour H_1 est de la forme

$$h_{1,\epsilon}\sigma_1^2 < (n-1)S_0^2 < h_{2,\epsilon}\sigma_1^2, \tag{8}$$

où $h_{i,\epsilon}$ se déterminent à partir des conditions (cf. condition (6.7) du théorème 6.1)

$$\mathbf{P}(h_{1,\epsilon} < \chi_{n-1}^2 < h_{2,\epsilon}) = 1 - \epsilon,$$
$$\mathbf{E}(\chi_{n-1}^2; h_{1,\epsilon} < \chi_{n-1}^2 < h_{2,\epsilon}) = (1 - \epsilon)\mathbf{E}\chi_{n-1}^2,$$
$$\chi_{n-1}^2 \in \mathbf{H}_{n-1}.$$

La région de confiance $\Theta^*(X, \epsilon)$ correspondant à (8) est l'intervalle

$$(n-1)S_0^2/h_{2,\epsilon} < \sigma^2 < (n-1)S_0^2/h_{1,\epsilon}. \tag{9}$$

Cet intervalle est visiblement invariant par g, de même d'ailleurs que le test (8) (dans cet exemple $G[\sigma_1] \equiv G$ pour tout σ_1). Donc, d'après les deuxièmes propositions des théorèmes 4 et 5, l'intervalle (9) est une région de confiance invariante sans biais la plus exacte au seuil $1 - \epsilon$.

EXEMPLE 3. Soit $X \in \Phi_{\alpha, \sigma^2}$. On demande de construire la région de confiance la plus exacte pour le paramètre α, σ étant inconnu. On a

$$f_{\alpha, \sigma^2}(X) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\mathrm{x}_i - \alpha)^2\right\}.$$

La famille Φ_{α, σ^2} sera invariante par le groupe G des transformations linéaires $gX = aX + b$ si l'on pose $\bar{g}(\alpha, \sigma) = (a\alpha + b, a\sigma)$. Le couple d'observations $(\bar{\mathrm{x}}, S_0^2)$ forme une statistique exhaustive. Il est aisé de voir qu'elle ne permet pas de construire une statistique invariante par G. Mais pour tout

α_1 on peut exhiber un sous-groupe $G[\alpha_1]$ des transformations $gX = a(X - \alpha_1) + \alpha_1$ par lequel la statistique $(\bar{x} - \alpha_1)/S_0$ sera un invariant maximal. L'hypothèse $H_1 = \{\alpha = \alpha_1\}$ reste invariante par $G[\alpha_1]$. En étudiant la densité de $(\bar{x} - \alpha_1)/S_0$ on démontre à l'aide des méthodes du § 7 (ces considérations seront omises pour leur complexité *) que pour tout σ il existe un test uniformément le plus puissant, invariant, sans biais de l'hypothèse H_1 contre l'hypothèse $\{\alpha \neq \alpha_1\}$ pour lequel la région d'acceptation de H_1 est

$$\sqrt{n}\,|\bar{x} - \alpha_1|/S_0 < \tau_\epsilon, \tag{10}$$

où τ_ϵ se détermine à partir de la condition $\mathbf{P}(|t_{n-1}| \geqslant \tau_\epsilon) = \epsilon$, $t_{n-1} \in \mathbf{T}_{n-1}$.

La région de confiance correspondante Θ^* est de la forme

$$\bar{x} - \tau_\epsilon S_0/\sqrt{n} < \alpha < \bar{x} + \tau_\epsilon S_0/\sqrt{n}. \tag{11}$$

Il est immédiat de voir que cet intervalle de confiance est invariant ($\Theta^*(gX, \epsilon) = \bar{g}\Theta^*(X, \epsilon)$). Le test (10) sera invariant par $G[\alpha_1]$ en vertu de la première proposition du théorème 5. L'intervalle (11) sera un intervalle de confiance sans biais invariant le plus exact (uniformément en σ) au seuil $1 - \epsilon$ en vertu de la deuxième proposition.

Nous avons donc établi dans ce paragraphe que tous les intervalles de confiance construits au § 2.32 étaient optimaux dans un certain sens.

§ 9. Approches bayésienne et minimax de test d'hypothèses multiples

1. Tests bayésiens et minimax. Les approches bayésienne et minimax ont été décrites au § 4. Les définitions nécessaires qui y ont été données seront rappelées au fil de l'exposé.

Supposons comme précédemment que l'on teste l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ contre $H_2 = \{\theta \in \Theta_2\}$ au vu d'un échantillon $X \in \mathbf{P}_\theta$.

L'approche totalement bayésienne implique que θ soit choisi au hasard avec une distribution *a priori* $\mathbf{Q}$ sur $\Theta = \Theta_1 \cup \Theta_2$. La distribution $\mathbf{Q}$ induit des distributions $\mathbf{Q}_i$ sur Θ_i, $i = 1, 2$, et des probabilités $q(i) = \mathbf{Q}(\theta \in \Theta_i)$, de sorte que $\mathbf{Q} = q(1)\mathbf{Q}_1 + q(2)\mathbf{Q}_2$. Désignons par $H_{\mathbf{Q}_i}$ l'hypothèse que $\theta \in \Theta_i$ est choisi au hasard avec la distribution $\mathbf{Q}_i$. D'après cette hypothèse la densité de X est

$$f_{\mathbf{Q}_i}(x) = \int f_\theta(x)\mathbf{Q}_i(dx).$$

Il est convenu de toute évidence (cf. § 4) que sur Θ_i sont définies des tribus $\mathfrak{C}_i$ sur lesquelles sont données $\mathbf{Q}_i$ et que $f_\theta(x)$ est mesurable par rapport à $\mathfrak{C}_i \times \mathfrak{B}^n_{\mathscr{X}}$.

---

*) Pour plus de détails cf. [50].

Des résultats des §§ 1 et 2 il s'ensuit que le test bayésien π_Q de H_{Q_1} contre H_{Q_2} utilisé dans le problème décrit ci-dessus sera de la forme

$$\pi_Q(X) = \begin{cases} 1 & \text{si} \quad f_{Q_2}(X) > cf_{Q_1}(X), \\ p & \text{si} \quad f_{Q_2}(X) = cf_{Q_1}(X), \\ 0 & \text{si} \quad f_{Q_2}(X) < cf_{Q_1}(X), \end{cases} \tag{1}$$

où $c = q(1)/q(2)$, et $p \in [0, 1]$ est arbitraire.

L'approche partiellement bayésienne est liée au test de l'hypothèse H_{Q_1} contre H_{Q_2} dans le cas où il n'existe pas de distribution *a priori* entre H_{Q_1} et H_{Q_2} (définie par les probabilités $q(1)$, $q(2)$). Posons

$$K_\epsilon^{Q_1} = \{\pi : \mathbf{E}_{Q_1}\pi(X) \leqslant \epsilon\}.$$

On dira alors qu'un test $\pi_{Q_1Q_2}$ est *bayésien* dans $K_\epsilon^{Q_1}$ s'il est le plus puissant de niveau $1 - \epsilon$ de H_{Q_1} contre H_{Q_2}. Le test $\pi_{Q_1Q_2}$ sera de la même forme (1), où c et p sont déduits de la condition $\mathbf{E}_{Q_1}\pi_{Q_1Q_2}(X) = \epsilon$.

On écrira π_{Q_1} ou π_{Q_2} au lieu de $\pi_{Q_1Q_2}$ si l'un des ensembles Θ_1 ou Θ_2 dégénère en un singleton $\{\theta_1\}$ ou $\{\theta_2\}$.

Dans les applications, on n'a pas souvent affaire à des problèmes dans lesquels les distributions $\mathbf{Q}_i$ sont entièrement connues. Mais comme nous l'avons vu à maintes reprises, l'intérêt de l'approche bayésienne ne se limite pas à la seule possibilité de son application directe. Cette approche permet de construire des tests uniformément les plus puissants ainsi que des tests minimax (comparer avec les §§ 1, 5, 6). Nous utiliserons plus loin l'approche bayésienne pour construire aussi des tests asymptotiquement optimaux. Supposons comme précédemment que

$$K_\epsilon = \{\pi : \sup_{\theta\in\Theta_1} \mathbf{E}_\theta\pi(X) \leqslant \epsilon\}. \tag{2}$$

On dit alors qu'un test $\bar{\pi}$ est *minimax* dans K_ϵ (resp. $K_\epsilon^{Q_1}$) si $\bar{\pi} \in K_\epsilon$ (resp. $\bar{\pi} \in K_\epsilon^{Q_1}$) et s'il maximise

$$\inf_{\theta\in\Theta_2} \mathbf{E}_\theta\pi(X) = \inf_{\theta\in\Theta_2} \beta(\theta). \tag{3}$$

Signalons que si les puissances $\beta(\theta) = \mathbf{E}_\theta\pi(X)$ sont continues et les ensembles Θ_1 et Θ_2 sont tangents, on a

$$\beta = \sup_{\pi\in K_\epsilon} \inf_{\theta\in\Theta_2} \beta(\theta) \leqslant \epsilon \tag{4}$$

et l'inégalité $\beta > \epsilon$ ne peut être réalisée. Si donc l'on désire que la puissance (3) soit suffisamment grande (tout au moins plus grande que ϵ), il faut envisager des ensembles Θ_1 et Θ_2 « séparés ». En d'autres termes, la région des

valeurs θ telles que $\beta(\theta)$ est voisine de ϵ doit être retirée en tant que région d'« indifférence » des tests et l'ensemble non tangent à Θ_1 pris pour Θ_2.

Mais si les ensembles Θ_1 et Θ_2 sont tangents, *tout test sans biais de* K_ϵ *sera minimax*. En effet, pour les tests sans biais on a $\beta(\theta) = \mathbf{E}_\theta \pi(X) \geqslant \epsilon$, $\theta \in \Theta_2$, et par suite, $\beta = \inf\limits_{\theta\in\Theta_2} \beta(\theta) \geqslant \epsilon$ atteint son maximum en vertu de (4).

La réciproque est vraie dans le cas général: *s'il existe un test minimax, il est sans biais*. Ceci résulte de ce que

$$\beta = \sup_{\pi\in \check{K}_\epsilon} \inf_{\theta\in\Theta_2} \beta(\theta) \geqslant \epsilon$$

(nous pouvons prendre $\pi(X) \equiv \epsilon$) et du fait que pour un test minimax on a

$$\inf_{\theta\in\Theta_2} \beta(\theta) = \beta.$$

Tout test uniformément le plus puissant sans biais $\check{\pi}$ *de la classe* $\check{K}_\epsilon$ *des tests sans biais est minimax dans* K_ϵ. En effet, soit $\check{\beta}(\theta)$ la puissance de π. Pour tous $\pi \in \check{K}_\epsilon$ et $\theta \in \Theta_2$, on a

$$\begin{gathered} \check{\beta}(\theta) \geqslant \beta(\theta), \ \inf_{\theta\in\Theta_2} \check{\beta}(\theta) \geqslant \inf_{\theta\in\Theta_2} \beta(\theta), \\ \inf_{\theta\in\Theta_2} \check{\beta}(\theta) = \sup_{\pi\in\check{K}_\epsilon} \inf_{\theta\in\Theta_2} (\theta) = \sup_{\pi\in\check{K}_\epsilon} \inf_{\theta\in\Theta_2} \beta(\theta). \end{gathered} \tag{5}$$

La dernière égalité s'explique par le fait que l'adjonction à $\check{K}_\epsilon$ des tests de K_ϵ pour lesquels $\inf\limits_{\theta\in\Theta_2} \beta(\theta) < \epsilon$ ne modifie pas la quantité $\sup\limits_{\pi\in\check{K}_\epsilon}$ dans (5). ◄

Dans le théorème 5.3 nous nous sommes servis des tests bayésiens pour chercher les tests uniformément les plus puissants. La proposition suivante « développe » en quelque sorte le théorème 5.3. Elle est analogue aussi aux théorèmes 1.2 et 2.11.2 et indique qu'il faut chercher les tests minimax dans la classe des tests (1) dont la forme est explicitement connue.

THÉORÈME 1. *Supposons qu'il existe des distributions* $\mathbf{Q}_i$ *concentrées respectivement sur des sous-ensembles* $\Theta_i^\circ \subset \Theta_i$, $i = 1, 2$, *et des constantes* c *et* p *telles que le test* $\pi_{\mathbf{Q}_1\mathbf{Q}_2}$ *défini dans* (1) *possède les propriétés*

1) $\pi_{\mathbf{Q}_1\mathbf{Q}_2} \in K_\epsilon^{\mathbf{Q}_1}$,

2) $$\mathbf{E}_\theta \pi_{\mathbf{Q}_1\mathbf{Q}_2}(X) = \sup_{\theta\in\Theta_1} \mathbf{E}_\theta \pi_{\mathbf{Q}_1\mathbf{Q}_2}(X) \tag{6}$$

pour tous les $\theta \in \Theta_1^\circ$,

3) $$\mathbf{E}_\theta \pi_{\mathbf{Q}_1\mathbf{Q}_2}(X) = \inf_{\theta\in\Theta_2} \mathbf{E}_\theta \pi_{\mathbf{Q}_1\mathbf{Q}_2}(X) \tag{7}$$

pour tous les $\theta \in \Theta_2^\circ$.

Alors $\pi_{Q_1Q_2} \in K_\epsilon$ *est un test minimax de* H_1 *contre* H_2.

Un couple de distributions Q_1, Q_2 *douées des propriétés* 2) *et* 3) *est le plus défavorable en ce sens que pour tout autre couple de distributions* Q_1', Q_2', *on a*

$$\inf_{\theta\in\Theta_2} E_\theta \pi_{Q_1Q_2} \leqslant \inf_{\theta\in\Theta_2} E_\theta \pi_{Q_1'Q_2'},$$

où $\pi_{Q_1'Q_2'}$ *est un test de* K_ϵ *de la forme* (1).

La dernière proposition exprime que de tous les tests bayésiens (1) le test $\pi_{Q_1Q_2}$ possède la plus petite puissance garantie.

DÉMONSTRATION. Puisque

$$\sup_{\theta\in\Theta_1} E_\theta \pi_{Q_1Q_2}(X) = \int_{\Theta_1^\circ} E_\theta \pi_{Q_1Q_2} Q_1(d\theta) = E_{Q_1}\pi_{Q_1Q_2} = \epsilon,$$

il vient que $\pi_{Q_1Q_2} \in K_\epsilon$. La puissance garantie de $\pi_{Q_1Q_2}$ est égale à (cf. (7))

$$\inf_{\theta\in\Theta_2} E_\theta \pi_{Q_1Q_2}(X) = \int_{\Theta_2^\circ} E_\theta \pi_{Q_1Q_2} Q_2(d\theta) = E_{Q_2}\pi_{Q_1Q_2} \equiv \beta_{Q_1Q_2}. \qquad (8)$$

Soit maintenant π un autre test dans K_ϵ de H_1 contre H_2. Alors π sera simultanément un test dans $K_\epsilon^{Q_1}$ de H_{Q_1} contre H_{Q_2}, car

$$E_{Q_1}\pi(X) = \int_{\Theta_1^\circ} E_\theta \pi(X) Q_1(d\theta) \leqslant \sup_{\theta\in\Theta_1} E_\theta \pi(X) \leqslant \epsilon. \qquad (9)$$

Mais $\pi_{Q_1Q_2}$ est le plus puissant de H_{Q_1} contre H_{Q_2} dans $K_\epsilon^{Q_1}$, donc, en vertu de (8)

$$\inf_{\theta\in\Theta_2} E_\theta \pi_{Q_1Q_2}(X) = \beta_{Q_1Q_2} \geqslant E_{Q_2}\pi(X) \geqslant \inf_{\theta\in\Theta_2} E_\theta \pi(X). \qquad (10)$$

Ce qui prouve la première proposition du théorème. Soient maintenant Q_1' et Q_2' deux autres distributions quelconques sur Θ_1 et Θ_2 respectivement. Le test $\pi_{Q_1Q_2}$ sera comme $\pi_{Q_1'Q_2'}$ un test dans $K_\epsilon^{Q_1'}$ de $H_{Q_1'}$ contre $H_{Q_2'}$, puisque

$$E_{Q_1'}\pi_{Q_1Q_2}(X) = \int_{\Theta_1} E_\theta \pi_{Q_1Q_2}(X) Q_1'(d\theta) \leqslant \sup_{\theta\in\Theta_1} E_\theta \pi_{Q_1Q_2}(X) \leqslant \epsilon.$$

Mais le test $\pi_{Q_1'Q_2'}$ est le plus puissant pour ces hypothèses, donc en vertu de (8)

$$\beta_{Q_1'Q_2'} \equiv E_{Q_2'}\pi_{Q_1'Q_2'}(X) \geqslant E_{Q_2'}\pi_{Q_1'Q_2'}(X) =$$
$$= \int_{\Theta_2} E_\theta \pi_{Q_1Q_2}(X) Q_2'(d\theta) \geqslant \inf_{\theta\in\Theta_2} E_\theta \pi_{Q_1Q_2}(X) \equiv \beta_{Q_1Q_2}. \blacktriangleleft$$

La principale difficulté soulevée par l'application du théorème 1 à des problèmes concrets est qu'il faut chercher (ou deviner) les distributions $\mathbf{Q}_1$ et $\mathbf{Q}_2$ les plus défavorables. L'invariance peut parfois nous être utile dans cette tâche comme nous le verrons dans les exemples du numéro suivant. Ces exemples présentent un intérêt en soi et seront utilisés dans la suite.

2. Tests minimax pour le paramètre α des distributions normales.

EXEMPLE 1. Soit $X = x_1 \in \Phi_{\alpha, E}$ un échantillon de taille $n = 1$ suivant une distribution normale à m dimensions de moyenne $\alpha = (\alpha_1, \ldots, \alpha_m)$ et de matrice des moments d'ordre deux unité. Posons $|\alpha|^2 = \sum_{i=1}^{m} \alpha_i^2$ et considérons le test de l'hypothèse $H_1 = \{|\alpha| \leqslant a\}$ contre $H_2 = \{|\alpha| \geqslant b\}, b > a$ (il existe ici une région « séparatrice » $a < |\alpha| < b$).

Si par exemple X représente dans un canal de transmission les amplitudes d'un vecteur-signal composé du « bruit » $X_0 \in \Phi_{0,1}$ et du signal utile α, $|\alpha| \geqslant b$, les hypothèses H_i pour $a = 0$ peuvent être alors considérées comme les hypothèses de la présence du signal utile.

Vu que l'exemple envisagé sera utilisé à maintes reprises dans la suite, on énoncera sous forme de théorème la proposition relative à la forme des tests minimax.

THÉORÈME 2. *Les tests minimax $\bar{\pi} \in K_\epsilon$ de $H_1 = \{|\alpha| \leqslant a\}$ contre $H_2 = \{|\alpha| \geqslant b\}, a < b$, au vu de $X \in \Phi_{\alpha, E}$ sont de la forme*

$$\bar{\pi}(X) = \begin{cases} 1 & si\ |X| > c_\epsilon, \\ 0 & sinon, \end{cases}$$

où c_ϵ est choisi à partir de la condition $p_c(a) = \epsilon$ et la puissance garantie de $\bar{\pi}$ est égale à $p_{c_\epsilon}(b)$

$$p_c(t) = \mathbf{P}((\xi_1 - t)^2 + \xi_2^2 + \ldots + \xi_m^2 > c^2),$$

ξ_i étant des variables aléatoires normales réduites indépendantes.

DÉMONSTRATION. Commençons par des raisonnements généraux. La densité de $x = (x^{(1)}, \ldots, x^{(m)})$ vaut ici

$$f_\alpha(x) = \frac{1}{(2\pi)^{m/2}} \exp\left\{-\frac{1}{2}(x - \alpha)(x - \alpha)^T\right\},$$

où x^T est un vecteur colonne. D'où il vient que la famille de distributions étudiée est invariante par la transformation orthogonale $gx = xC$, où C est la matrice d'une transformation orthogonale de R^m. Ceci étant, il faut poser $\bar{g}\alpha = \alpha C$. Les hypothèses H_i seront invariantes par $\bar{g}$.

Supposons par souci de simplicité que $a = 0$. Si la distribution $\mathbf{Q}_2$ sur

$\Theta_2 = \{\alpha : |\alpha| \geqslant b\}$ n'était pas invariante par $\bar{g}$ (ce cas se présenterait si par exemple elle était concentrée dans le voisinage d'un point quelconque α_0), cette absence de symétrie aurait pu être utilisée lors de la résolution du problème (sous la condition que nous venons juste de poser nous aurions eu presque affaire à un problème de test de deux hypothèses simples $\{\alpha = 0\}$ et $\{\alpha = \alpha_0\}$ et nous aurions obtenu un test d'une grande puissance). Donc, une telle distribution ne peut être la plus défavorable. Pour l'être, il faut qu'elle soit invariante par $\bar{g}$. Par ailleurs, il est clair que la pire des situations est celle où toute la distribution est concentrée sur la frontière de Θ_2 (plus des hypothèses sont voisines et plus il est difficile de les distinguer). On peut se livrer aux mêmes raisonnements généraux sur $\mathbf{Q}_1$ si $a \neq 0$.

Il semble donc naturel que les distributions $\mathbf{Q}_1$ et $\mathbf{Q}_2$ les plus défavorables dans notre exemple soient les distributions uniformes sur les sphères $\Theta_1^\circ = \{\alpha : |\alpha| = a\}$ et $\Theta_2^\circ = \{\alpha : |\alpha| = b\}$. Dans ce cas, le théorème 1 affirme qu'un test minimax $\bar{\pi}$ sera de la forme $\bar{\pi}(x) = \pi_{\mathbf{Q}_1\mathbf{Q}_2}(x)$, où $\pi_{\mathbf{Q}_1\mathbf{Q}_2}(x) = 1$ si

$$\int_{\Theta_2^\circ} \exp\left\{-\frac{1}{2}(x - v)(x - v)^T\right\}\frac{dV(v)}{V_2} >$$

$$> c\int_{\Theta_1^\circ} \exp\left\{-\frac{1}{2}(x - v)(x - v)^T\right\}\frac{dV(v)}{V_1} \qquad (11)$$

et $\pi_{\mathbf{Q}_1\mathbf{Q}_2}(x) = 0$ sinon. Ici $dV(v)$ désigne l'aire élémentaire de la sphère correspondante, $V_i =$ mes Θ_i°, $i = 1, 2$.

Considérons n'importe laquelle de ces intégrales, celle de droite par exemple, et remarquons qu'on peut la mettre sous la forme

$$\exp\left\{-\frac{1}{2}xx^T - a^2\right\}\cdot \int_{\Theta_1^\circ} \exp\{xv^T\}\,\frac{dV(v)}{V_1}\,.$$

L'intégrale est ici égale à

$$\int_{\Theta^\circ} \exp\{|x|ae_x v^T\}dV(v)/V, \qquad V = \text{mes}\,\Theta^\circ,$$

où Θ° est la sphère unité, $e_x = x/|x|$. Si donc l'on pose

$$\psi(t) = \int_{\Theta^\circ} \exp\{te_x v^T\}dV(v), \qquad (12)$$

la région d'acceptation (11) de H_2 devient

$$\psi(|x|b) > c\psi(|x|a) \tag{13}$$

(par c on désigne des constantes qui ne sont pas nécessairement confondues avec celles de (11)). Or $\psi(t)$ ne dépend visiblement pas de x, puisque l'intégrale (12) ne dépend pas du sens du vecteur e_x. Donc

$$\psi(t) = \int_{\Theta^\circ} \exp\{tv_1\} dV(v),$$

où v_1 est la première composante du vecteur v.

La fonction $\psi(t)$ est convexe et strictement croissante sur $[0, \infty[$ puisque $\psi'(0) = 0$ et $\psi''(t) > 0$ pour $t > 0$. Il s'ensuit que l'inégalité (13) (ou (11)) est équivalente à

$$|x| > c. \tag{14}$$

On reconnaît visiblement un test invariant. Assurons-nous qu'il remplit bien les conditions 1, 2 et 3 du théorème 1, ce qui exprimera qu'il est minimax.

On a

$$\mathbf{E}_\alpha \pi_{Q_1Q_2}(X) = \mathbf{P}_\alpha(|X| > c) = \Phi_{0,E}(\{x : |x - \alpha| > c\}).$$

Il est clair que cette probabilité ne change pas lorsque le point α se déplace sur la sphère $|\alpha| = \text{const}$. Elle ne dépend donc que de $|\alpha|$ et par suite

$$\mathbf{E}_\alpha \pi_{Q_1Q_2} = \mathbf{P}(|\xi - \alpha|^2 > c^2) =$$

$$= \mathbf{P}\left(\sum_{i=1}^{m} (\xi_i - \alpha_i)^2 > c^2\right) = \mathbf{P}((\xi_1 - |\alpha|)^2 + \xi_2^2 + \ldots + \xi_m^2 > c^2),$$

où $\xi_i \in \Phi_{0,1}$ sont les composantes indépendantes du vecteur ξ.

LEMME 1. *La fonction* $p_c(t) = \mathbf{P}((\xi_1 - t)^2 + \xi_2^2 + \ldots + \xi_m^2 > c^2)$ *est une fonction de* $|t|$ *strictement croissante pour tout* c.

Ce lemme entraîne

$$\mathbf{E}_\alpha \pi_{Q_1Q_2}(X) = p_c(|\alpha|) \leqslant p_c(a) \quad \text{si} \quad |\alpha| \leqslant a,$$
$$\mathbf{E}_\alpha \pi_{Q_1Q_2}(X) = p_c(|\alpha|) \geqslant p_c(b) \quad \text{si} \quad |\alpha| \geqslant b.$$

Ces relations sont équivalentes aux conditions 2) et 3) du théorème 1. Pour que le test $\pi_{Q_1Q_2}$ soit un test de niveau $1 - \epsilon$, nous devons poser c égal à la solution c_ϵ de l'équation $p_c(a) = \epsilon$. Le test $\pi_{Q_1Q_2}$ est donc un test minimax de niveau $1 - \epsilon$ et sa puissance garantie est égale à $p_{c_\epsilon}(b)$. ◄

DÉMONSTRATION du lemme 1. On peut se limiter aux valeurs $t \geqslant 0$, puisque $p_c(t) = p_c(-t)$.

Traitons d'abord le cas où $m = 1$. Désignons la fonction $p_c(t)$ par $p(t)$. On a

$$p(t) = \mathbf{P}(|\xi_1 - t|^2 > c^2) = \Phi(t - c) + 1 - \Phi(t + c).$$

Donc, la dérivée par rapport à t est égale à

$$p'(t) = \frac{1}{\sqrt{2\pi}} [e^{-(t-c)^2/2} - e^{-(c+t)^2/2}] =$$
$$= \frac{1}{\sqrt{2\pi}} e^{-(c^2+t^2)/2} [e^{ct} - e^{-ct}] \geqslant 0,$$

et la fonction $p(t)$ est strictement croissante pour $t \geqslant 0$.

Pour $m > 1$ la fonction $p_c(t)$ est le produit de convolution de la fonction $p(t) = p(t, c^2)$ et de la distribution du χ^2 à $m - 1$ degrés de liberté :

$$p_c(t) = \int_0^\infty p(t, c^2 - u) d\mathbf{H}_{m-1}(u).$$

Il est évident que c'est aussi une fonction strictement croissante de t pour $t \geqslant 0$. ◁

Faisons la remarque suivante sur le théorème 2. Supposons pour simplifier que $a = 0$. L'hypothèse $H_1 = \{\alpha = 0\}$ devient alors simple. Si l'on construit un test le plus puissant pour chaque contre-hypothèse $\alpha \in \Theta_2$, on obtient un test de la forme

$$x\alpha^T > c.$$

Ceci exprime que chaque direction $\alpha = \alpha_0 t$, $\alpha_0 \in \Theta_2^\circ$, $t \geqslant 1$, possédera son propre test le plus puissant de niveau $1 - \epsilon$

$$x\alpha_0^T > c_\epsilon, \tag{15}$$

où c_ϵ dépend uniquement de ϵ, puisque $\mathbf{E}_0(X\alpha_0^T) = 0$, $\mathbf{V}_0(X\alpha_0^T) = |\alpha_0|^2 = b$. Mais la région critique d'un test minimax (invariant) doit être indifféremment sensible à toutes les contre-hypothèses. C'est pourquoi cette région est la réunion de demi-espaces (15) qui a la forme de l'extérieur d'une sphère.

EXEMPLE 2. Supposons maintenant que $X = x_1 \in \Phi_{\alpha, \sigma^2}$, où $\sigma^2 = \|\sigma_{ij}\|$ est une matrice définie positive des moments d'ordre deux. Soit à tester l'hypothèse $H_1 = \{\alpha\sigma^{-2}\alpha^T \leqslant a^2\} = \{|\alpha\sigma^{-1}| \leqslant a\}$ contre l'hypothèse $H_2 = \{\alpha\sigma^{-2}\alpha^T \geqslant b^2\} = \{|\alpha\sigma^{-1}| \geqslant b\}$, $a < b$. Le théorème 2 entraîne le

THÉORÈME 2A. *La région critique d'un test minimax de niveau* $1 - \epsilon$ *de* H_1 *contre* H_2 *est de la forme*

$$x\sigma^{-2}x^T > c_\epsilon^2$$

et la puissance garantie est égale à $p_{c_\epsilon}(b)$, où c_ϵ est comme précédemment solution de l'équation $p_c(a) = \epsilon$.

DÉMONSTRATION. Posons $gx = x\sigma$ et remarquons qu'en vertu de (7.3)

$$\Phi_{\alpha, E}(A) = \Phi_{\bar{g}(\alpha, E)}(gA),$$

où $\bar{g}(\alpha, E) = (\alpha\sigma, \sigma^2)$. Pour la boule $A = \{x : |x| < c\}$ on aura

$$\begin{aligned} gA &= \{y = x\sigma : xx^T < c^2\} = \{y : y\sigma^{-2}y^T < c^2\}, \\ \Phi_{\alpha, E}(A) &= \Phi_{\alpha\sigma, \sigma^2}(\{x : x\sigma^{-2}x^T < c^2\}). \end{aligned} \tag{16}$$

L'image de l'ensemble $\{\alpha : |\alpha| \leqslant a\}$ par l'application $\bar{g}$ est l'ensemble $\{\beta = \alpha\sigma : \alpha\alpha^T \leqslant a^2\} = \{\beta : \beta\sigma^{-2}\beta^T \leqslant a^2\}$.

Donc, toutes les relations établies dans l'exemple 1 pour $\Phi_{\alpha, E}(A)$ dans $|\alpha| \leqslant a$ ou $|\alpha| \geqslant b$ seront valables pour $\Phi_{\beta, \sigma^2}(\{x : x\sigma^{-2}x^T < c^2\})$ dans $|\beta\sigma^{-1}| \leqslant a$ ou $|\beta\sigma^{-1}| \geqslant b$ respectivement.

Ce qui prouve le théorème 2A. ◁

EXEMPLE 3. Considérons de nouveau un échantillon issu d'une distribution normale $\Phi_{\alpha, E}$ et de matrice des moments d'orde deux unité. Mais contrairement à l'exemple 1 les hypothèses H_i ne porteront que sur une partie des coordonnées du vecteur α. Représentons α sous la forme de deux vecteurs $\alpha = (\alpha', \alpha'')$, où $\alpha' = (\alpha_1, ..., \alpha_l)$ et $\alpha'' = (\alpha_{l+1}, ..., \alpha_m)$ et soit à tester l'hypothèse $H_1 = \{|\alpha''| \leqslant a\}$ contre $H_2 = \{|\alpha''| \geqslant b\}$ au vu de l'échantillon $X = \mathrm{x}_1 = (\mathrm{x}_{1,1}, ..., \mathrm{x}_{1,m})$ de taille $n = 1$. Dans chacune de ces hypothèses, la quantité α' peut prendre une valeur arbitraire. Procédons exactement comme dans l'exemple 1, mais prenons pour $\mathbf{Q}_1$ et $\mathbf{Q}_2$ des distributions uniformes sur les « sphères » $\Theta_1^0 = \{\alpha : |\alpha''| = a, \alpha' = \alpha_0'\}$, $\Theta_2^0 = \{\alpha : |\alpha''| = b, \alpha' = \alpha_0'\}$, où α_0' est un point quelconque donné. Si l'on pose $\mathrm{x}_1' = (\mathrm{x}_{1,1}, ..., \mathrm{x}_{1,l})$, $\mathrm{x}_1'' = (\mathrm{x}_{1,l+1}, ..., \mathrm{x}_{1,m})$, on obtient en définitive un test minimax

$$|\mathrm{x}_1''| > c_\epsilon,$$

où c_ϵ est solution de l'équation

$$\mathbf{P}((\xi_1 - a)^2 + \xi_2^2 + ... + \xi_{m-l}^2 > c^2) = \epsilon \tag{17}$$

(les facteurs $\exp\left\{-\dfrac{1}{2}(x' - \alpha_0')(x' - \alpha_0')^T\right\}$ se simplifient dans l'inégalité $f_{\mathbf{Q}_2}(X)/f_{\mathbf{Q}_1}(X) > c$ et celle-ci se transforme en une inégalité de type (11)). Ce résultat est tout à fait naturel, puisque les coordonnées $\mathrm{x}_{1,j}$ sont indépendantes ici, et par suite, le sous-vecteur x_1' ne contient aucune information sur α''. Donc, de l'échantillon $X = \mathrm{x}_1$ il ne suffit de considérer que le sous-vecteur x_1'', et le problème se ramène alors à l'exemple 1.

Le test des hypothèses de l'exemple 3 fait partie des problèmes mettant

en jeu un paramètre « fantôme ». Dans notre cas il s'agit du vecteur α'. Pour les raisons indiquées ci-dessus ce vecteur ne nous a en fait pratiquement pas empêché de construire un test minimax qui est *ipso facto* indépendant de α'.

La situation est différente dans l'exemple suivant qui traite du cas plus général où les coordonnées x_{1j} sont dépendantes.

EXEMPLE 4. Soit $X = x_1 \in \Phi_{\alpha,\sigma^2}$. Soit à tester l'hypothèse

$$H_1 = \{\alpha d^{-2}\alpha^T \leqslant a^2\} \quad \textit{contre} \quad H_2 = \{\alpha d^{-2}\alpha^T \geqslant b^2\}, \tag{18}$$

où d^{-2} est une matrice semi-définie positive de rang $m - l < m$, obtenue à partir de σ^{-2} en remplaçant par des zéros les éléments de l lignes quelconques et des l colonnes de mêmes numéros. Pour fixer les idées, on peut admettre que pour la matrice définie positive σ_2^{-2} d'ordre $m - l$, inverse de la matrice

$$\sigma_2^2 = \mathbf{E}_{\alpha,\sigma^2}(x_1'' - \alpha'')^T(x_1'' - \alpha''),$$

formée par les $m - l$ dernières lignes et colonnes de la matrice $\sigma^2 = \|\sigma_{ij}\|$, on teste l'hypothèse $H_1 = \{\alpha''\sigma_2^{-2}\alpha''^T \leqslant a^2\}$ contre $H_2 = \{\alpha''\sigma_2^{-2}\alpha''^T \geqslant b^2\}$, où x_1'' et α'' désignent les mêmes sous-vecteurs de x_1 et α que dans l'exemple précédent. Le paramètre α' peut être arbitraire dans chacune des hypothèses H_i.

Dans cet exemple la distribution de x_1' dépend généralement de α''. Orthonormons le vecteur x_1. Posons

$$y = x_1\Lambda, \tag{19}$$

où $\Lambda = \|a_{ij}\|$ est une matrice triangulaire dont les éléments $a_{ij} = 0$ pour $j > i$, et les autres sont choisis à partir de la condition $y \in \Phi_{\beta,E}$, où $\beta = (\beta_1, \ldots, \beta_m) = \alpha\Lambda$. Ceci est toujours possible, puisque de (19) il vient

$$y_m = x_{1,m}a_{m,m},$$
$$y_{m-1} = x_{1,m}a_{m,m-1} + x_{1,m-1}a_{m-1,m-1},$$
$$\dots\dots\dots\dots\dots\dots\dots\dots$$

De là et des conditions

$$\mathbf{E}_{\alpha,\sigma^2}(y_i - \beta_i)^2 = 1,$$
$$\mathbf{E}_{\alpha,\sigma^2}(y_i - \beta_i)(y_j - \beta_j) = 0, \quad i \neq j,$$

on obtient successivement les valeurs

$$a_{m,m}^2 = 1/\sigma_{m,m},$$
$$\sigma_{m,m}a_{m,m-1} + \sigma_{m-1,m}a_{m-1,m-1} = 0,$$
$$\sigma_{m,m}a_{m,m-1}^2 + 2\sigma_{m,m-1}a_{m,m-1}a_{m-1,m-1} + \sigma_{m-1,m-1}a_{m-1,m-1}^2 = 1,$$
$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots$$

La matrice triangulaire Λ est donc telle que

$$\mathbf{E}_{\alpha,\sigma^2}(y-\beta)^T(y-\beta) = \mathbf{E}_{\alpha,\sigma^2}\Lambda^T(x_1-\alpha)^T(x_1-\alpha)\Lambda = \Lambda^T\sigma^2\Lambda = E.$$

La triangularité de Λ entraîne que le vecteur $\beta'' = (\beta_{l+1}, \ldots, \beta_m)$ ne dépend que de α'', et réciproquement. Si l'on désigne par Λ_2 la matrice triangulaire d'ordre $m-l$ formée des $m-l$ dernières lignes et colonnes de Λ, on trouve de toute évidence que $\beta'' = \alpha''\Lambda_2$, $\Lambda_2^T\sigma_2^2\Lambda_2 = E$. L'image de l'ensemble $\Theta_1 = \{\alpha : \alpha''\sigma_2^{-2}\alpha''^T \leqslant a^2\}$ est

$$\{\beta : \beta = \alpha\Lambda, \alpha''\sigma_2^{-2}\alpha''^T \leqslant a^2\} = \{\beta : \beta''\Lambda_2^{-1}\sigma_2^{-2}\Lambda_2^{-1T}\beta''^T \leqslant a^2\} =$$
$$= \{\beta : \beta''\beta''^T \leqslant a^2\} = \{\beta : |\beta''| \leqslant a\}.$$

Le « sous-paramètre » β' peut être arbitraire si α' l'est.

Nous sommes arrivés au problème de l'exemple 3. Un test minimax de niveau $1-\epsilon$ de H_1 contre H_2 est par conséquent de la forme $y''y''^T > c_\epsilon$ ou ($\Lambda_2\Lambda_2^T = \sigma_2^{-2}$)

$$x_1''\sigma_2^{-2}x_1''^T > c_\epsilon,$$

où c_ϵ est solution de l'équation (17).

Le dernier exemple est le plus général des exemples 1 à 4. Il les résume de la manière suivante.

THÉORÈME 2B. *Si au vu d'un échantillon $X = x_1 \in \Phi_{\alpha,\sigma^2}$ on teste les hypothèses* (18) *liées à la valeur $\alpha d^{-2}\alpha^T$, un test minimax de niveau $1-\epsilon$ est de la forme*

$$x_1 d^{-2} x_1^T > c_\epsilon, \tag{20}$$

où c_ϵ se déduit de (17), *$m-l$ étant le rang de d^{-2}.*

La puissance garantie du test (20) *est égale à*

$$\mathbf{P}((\xi_1-b)^2 + \xi_2 + \ldots + \xi_{m-l} > c_\epsilon^2), \quad \xi_i \in \Phi_{0,1}.$$

Si l'échantillon X est de taille n, alors $\bar{x} \in \Phi_{\alpha,\sigma^2/n}$ sera une statistique exhaustive et un test minimax sera de la forme

$$\bar{x}d^{-2}\bar{x}^T > c_\epsilon/n.$$

L'exemple suivant est de nature légèrement différente.

EXEMPLE 5. Supposons comme dans l'exemple 1 que $X = x_1 \in \Phi_{\alpha,E}$ est un échantillon de taille $n = 1$ issu d'une distribution normale m-dimensionnelle de moyenne $\alpha = (\alpha_1, \ldots, \alpha_m)$. Supposons que $H_1 = \{\alpha = 0\}$ et que H_2 consiste en ce que α appartient à un ensemble $\bar{\Theta}_2$ ne contenant pas $\alpha \in \Theta_2$. Désignons par $\bar{\Theta}_2$ l'adhérence convexe de Θ_2 (le plus petit fermé convexe contenant Θ_2) et soit β le point de $\bar{\Theta}_2$ le plus proche de l'origine des coordonnées. Si $\beta \in \Theta_2$, la distribution $\mathbf{Q}_2$ concentrée au point β sera la plus

défavorable et $\bar{\pi}$ sera de la forme $\bar{\pi}(X) = 1$ si

$$(X - \beta)(X - \beta)^T < XX^T + c_1$$

ou, ce qui revient au même, si

$$X\beta^T/|\beta| > c_2,$$

où c_2 se détermine à partir de la condition $\bar{\pi} \in K_\epsilon$.

En effet, il suffit de vérifier la condition (7). On a

$$\mathbf{E}_\alpha \bar{\pi}(X) = \mathbf{P}_\alpha(X\beta^T/|\beta| > c_2),$$

où $X\beta^T/|\beta| \in \Phi_{\alpha\beta^T/|\beta|,\,1}$, de sorte que

$$\mathbf{E}_\alpha \bar{\pi}(X) = 1 - \Phi(c_2 - \alpha\beta^T/|\beta|).$$

Ceci signifie que $\mathbf{E}_\alpha \bar{\pi}(X)$, $\alpha \in \Theta_2$, est minimisée par la valeur α qui minimise la fonction $\alpha\beta^T/|\beta|$. Mais il est évident que $\alpha\beta^T \geqslant \beta\beta^T = |\beta|^2$ pour tous les $\alpha \in \Theta_2$, de sorte que

$$\mathbf{E}_\beta \bar{\pi}(X) = \inf_{\alpha \in \Theta_2} \mathbf{E}_\alpha \bar{\pi}(X). \triangleleft$$

Nous proposons au lecteur de construire un test minimax dans le même problème pour le cas où $X \in \Phi_{\alpha,\sigma^2}$ et σ^2 est une matrice des moments d'ordre deux quelconque.

3. Distributions dégénérées les plus défavorables pour hypothèses unilatérales. Soit $X \in \mathbf{P}_\theta$, où θ et les éléments x_i de l'échantillon X sont réels.

Soit à tester une hypothèse unilatérale $H_1 = \{\theta \leqslant \theta_1\}$ contre $H_2 = \{\theta \geqslant \theta_2\}$ dans le cas d'une région d'« indifférence » $\theta_1 < \theta < \theta_2$ non vide. Il serait intéressant de savoir sous quelles conditions les distributions les plus défavorables seront concentrées aux points θ_1 et θ_2. En effet, le test minimax $\bar{\pi}$ de niveau $1 - \epsilon$ serait alors de la forme très simple

$$\bar{\pi}(X) = \begin{cases} 1 & \text{si} \quad f_{\theta_2}(X) > cf_{\theta_1}(X), \\ p & \text{si} \quad f_{\theta_2}(X) = cf_{\theta_1}(X), \\ 0 & \text{si} \quad f_{\theta_2}(X) < cf_{\theta_1}(X), \end{cases} \tag{21}$$

où p et c se définissent à partir de l'égalité $\mathbf{E}_{\theta_1} \bar{\pi}(X) = \epsilon$.

Nous savons déjà que si le rapport de vraisemblance est monotone, un tel test sera uniformément le plus puissant et par suite minimax. La proposition suivante nous fournit une autre condition suffisante pour qu'un test soit minimax.

THÉORÈME 3. *Supposons que la densité $f_\theta(x)$ est telle que le rapport $f_{\theta'}(x)/f_\theta(x)$ ne décroît pas par rapport à x pour tout $\theta' > \theta$. Alors les distributions les plus défavorables $\mathbf{Q}_1$ et $\mathbf{Q}_2$ seront concentrées respectivement aux points θ_1 et θ_2, et par suite, le test* (21) *sera minimax.*

DÉMONSTRATION. Supposons tout d'abord que $n = 1$. D'après l'hypothèse du théorème, il existe des $a \leqslant b$ tels que

$$f_{\theta'}(x)/f_\theta(x) \begin{cases} \leqslant 1 & \text{pour} \quad x \in]-\infty, a], \\ = 1 & \text{pour} \quad x \in]a, b[, \\ \geqslant 1 & \text{pour} \quad x \in [b, \infty[. \end{cases}$$

Comme $\bar{\pi}(x)$ est croissante, on a $\bar{\pi}(b) \geqslant \bar{\pi}(a)$ et

$$\mathbf{E}_{\theta'} \bar{\pi}(X) - \mathbf{E}_\theta \bar{\pi}(X) \geqslant$$

$$\geqslant \bar{\pi}(a) \int_{-\infty}^{a} (f_{\theta'}(x) - f_\theta(x))\mu(dx) + \bar{\pi}(b) \int_{b}^{\infty} (f_{\theta'}(x) - f_\theta(x))\mu(dx) =$$

$$= (\bar{\pi}(b) - \bar{\pi}(a)) \int_{b}^{\infty} (f_{\theta'}(x) - f_\theta(x))\mu(dx) \geqslant 0.$$

Si $n > 1$, pour obtenir une inégalité analogue, il faut se servir d'une intégration successive (d'abord par rapport à x_1, puis par rapport à x_2, et ainsi de suite) et du fait que $\bar{\pi}(X)$ est croissante par rapport à chacun de ses arguments.

Nous avons ainsi établi que la puissance $\beta(\theta) = \mathbf{E}_\theta \bar{\pi}(X)$ est une fonction croissante.

Il s'ensuit que le niveau de π est $1 - \epsilon$ et que $\beta(\theta_1) = \sup_{\theta \leqslant \theta_1} \beta(\theta), \beta(\theta_2) = \inf_{\theta \geqslant \theta_2} \beta(\theta)$. Ce qui exprime que toutes les conditions du théorème 1 sont remplies. Le théorème 3 est prouvé. ◄

Si θ est un paramètre de translation : $f_\theta(x) = f(x - \theta)$, on démontre que $f_{\theta'}(x)/f_\theta(x)$ sera monotone par rapport à x si et seulement si la fonction $-\ln f(x)$ est convexe (cf. [50]).

§ 10. Test du rapport de vraisemblance

Dans les paragraphes précédents nous avons acquis de nombreux résultats sur la construction de divers tests optimaux. La conclusion importante que l'on peut tirer est que ces tests optimaux n'existent que sous des conditions très restrictives. La situation était à peu de chose près la même en théorie de l'estimation : les estimateurs efficaces n'existaient que sous des conditions contraignantes. Mais nous avons vu au chapitre 2 que les estimateurs possédant la propriété d'efficacité asymptotique, existent assez souvent, sous des conditions assez larges liées essentiellement à la régularité de la famille $\{\mathbf{P}_\theta\}$. Tel est le cas des estimateurs du maximum de vraisemblance.

L'autre expression de l'optimalité asymptotique des estimateurs du maximum de vraisemblance consiste, comme nous l'avons vu, en ce que ces estimateurs sont asymptotiquement équivalents aux estimateurs bayésiens pour toute distribution *a priori* régulière donnée.

En théorie de test d'hypothèses, le *test du rapport de vraisemblance* est un peu analogue à l'estimateur du maximum de vraisemblance. Sous des hypothèses larges il est confondu avec les tests optimaux, si ceux-ci existent, et est asymptotiquement équivalent à un test bayésien dans le cas où $\Theta_1 = \{\theta_1\}$ pour toute distribution *a priori* régulière donnée $\mathbf{Q}_2$ sur Θ_2. Dans les prochains paragraphes nous établirons cette propriété ainsi que de nombreuses autres propriétés asymptotiques du test du rapport de vraisemblance.

Donnons la définition du test du rapport de vraisemblance. Dans le cas paramétrique où $X \in \mathbf{P}_\theta$, soit à tester l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ contre l'hypothèse $H_2 = \{\theta \in \Theta_2\}$.

DÉFINITION 1. On appelle *test du rapport de vraisemblance* de H_1 contre H_2 un test $\hat{\pi}(X)$ de région critique

$$R(X) \equiv \frac{\sup\limits_{\theta\in\Theta_2} f_\theta(X)}{\sup\limits_{\theta\in\Theta_1} f_\theta(X)} > c. \tag{1}$$

La constante c est généralement déterminée à partir de la condition

$$\sup_{\theta\in\Theta_1} \mathbf{P}_\theta(R(X) > c) = \epsilon, \tag{2}$$

condition sous laquelle le test du rapport de vraisemblance aura un niveau égal à $1 - \epsilon$.

Parallèlement au test (1) on envisage souvent un test de nature équivalente (appelé aussi test du rapport de vraisemblance) de la forme

$$R_1(X) \equiv \frac{\sup\limits_{\theta\in\Theta} f_\theta(X)}{\sup\limits_{\theta\in\Theta_1} f_\theta(X)} = \frac{f_{\theta^*}(X)}{\sup\limits_{\theta\in\Theta_1} f_\theta(X)} > c. \tag{3}$$

La proximité de ces tests résulte du fait que pour $\Theta = \Theta_1 \cup \Theta_2$,

$$f_{\theta^*}(X) = \max\left\{\sup_{\theta\in\Theta_1} f_\theta(X), \sup_{\theta\in\Theta_2} f_\theta(X)\right\},$$

et par suite $R_1(X) = \max\{1, R(X)\}$.

Si l'hypothèse H_1 est simple : $\Theta_1 = \{\theta_1\}$, $H_2 = \{\theta \neq \theta_1\}$, de sorte que $\Theta_2 = \Theta \setminus \{\theta_1\}$, alors pour les $f_\theta(x)$ continues par rapport à θ, on aura

$$R(X) = R_1(X) = f_{\theta^*}(X)/f_{\theta_1}(X).$$

De par sa forme le test (1) généralise de façon naturelle le test le plus puissant d'hypothèses simples du lemme de Neyman-Pearson. Bien que ce test ne possède probablement pas de propriétés d'optimalité *exactes* dans le cas général, il est souvent le meilleur asymptotiquement (cf. §§ 13 à 16).

De nombreux tests minimax et invariants sans biais, étudiés plus haut sont des tests du rapport de vraisemblance. Considérons à titre d'illustration les exemples 9.1 à 9.4 dans lesquels on a construit des tests minimax pour le paramètre α de distributions normales. *Dans tous ces exemples, les tests minimax étaient des tests du rapport de vraisemblance*. Prouvons-le. Les problèmes des exemples 9.2 et 9.4 ont été réduits, aux transformations linéaires près du paramètre, aux problèmes des exemples 9.1 et 9.3. Vu que le rapport de vraisemblance (1) n'est pas affecté par ces changements (les régions Θ_i étant modifiées en conséquence), il nous suffit de nous pencher seulement sur les exemples 9.1 et 9.3.

Dans l'exemple 9.1 on a testé l'hypothèse $H_1 = \{|\alpha| \leqslant a\}$ contre $H_2 = \{|\alpha| \geqslant b\}$, $a < b$, au vu d'un échantillon $X \in \Phi_{\alpha, E}$ de taille $n = 1$ issu d'une distribution normale multidimensionnelle dont la matrice des moments d'ordre deux est une matrice unité. Un test minimax est de la forme

$$|X| > c. \tag{4}$$

Dans ce cas, $\sup_{\theta \in \Theta_i} f_\theta(X)$ est défini par la valeur

$$\inf_{\alpha \in \Theta_i} (X - \alpha)(X - \alpha)^T = \inf_{\alpha \in \Theta_i} |X - \alpha|^2,$$

de sorte que pour la statistique $R(X)$ de (1), on aura

$$\ln R(X) = \begin{cases} -\frac{1}{2}(|X| - b)^2 & \text{si} \quad |X| \leqslant a, \\ -\frac{1}{2}(|X| - b)^2 + \frac{1}{2}(|X| - a)^2 & \text{si} \quad a < |X| < b, \\ -\frac{1}{2}(|X| - a)^2 & \text{si} \quad |X| \geqslant b. \end{cases}$$

Cette fonction est une fonction de $|X|$ continue strictement croissante. Donc, les régions (1) et (4) sont confondues pour des valeurs convenables de c.

On propose au lecteur de s'assurer que le test (3) est aussi de la forme (4) dans cet exemple.

Dans l'exemple 9.3, on a éprouvé l'hypothèse $H_1 = \{|\alpha''| \leqslant a\}$ contre $H_2 = \{|\alpha''| \geqslant b\}$, où $\alpha'' = (\alpha_{l+1}, \ldots, \alpha_m)$ est le sous-vecteur de α, com-

posé des $m - l$ dernières coordonnées, au vu d'un échantillon $X \in \Phi_{\alpha, E}$ de taille un. Un test minimax est de la forme

$$|X''| > c, \tag{5}$$

où X'' est composé des $m - l$ dernières coordonnées du vecteur X. Mais dans ce cas

$$\inf_{\alpha \in \Theta_1} (X - \alpha)(X - \alpha)^T = \inf_{\alpha'' : |\alpha''| \leqslant a} (X'' - \alpha'')(X'' - \alpha'')^T.$$

On a une inégalité de même nature pour Θ_2. Tout se ramène donc à l'exemple 9.1, et les tests du rapport de vraisemblance (1) et (3) seront confondus avec (5).

Dans les conditions du § 5 les tests uniformément les plus puissants pour les familles exponentielles

$$f_\theta(x) = c(\theta) e^{\theta T(x)} h(x) \tag{6}$$

seront également confondus avec un test du rapport de vraisemblance. Le lecteur peut s'en assurer en remarquant que la fonction

$$\varphi(\theta) = \ln c(\theta) = -\ln \left(\int e^{\theta T(x)} h(x) \mu^n(dx) \right)$$

est convexe, puisque $\varphi'(\theta) = -\mathbf{E}_\theta T$, $\varphi''(\theta) = -\mathbf{V}_\theta T < 0$. De la convexité de φ il s'ensuit que l'équation

$$\varphi'(\theta) + T(X) = 0$$

admet une solution unique pour l'estimateur du maximum de vraisemblance $\hat{\theta}^* = \psi(T)$ et que la fonction ψ est monotone. Ceci étant, l'un des $\sup_{\theta \in \Theta_i} f_\theta(X)$ sera atteint au point $\hat{\theta}^*$, l'autre, en θ_1 ou θ_2.

La vérification de cette assertion pour les familles normales $\Phi_{\alpha, E}$ qui sont un cas particulier de (6) est accessible dans le § 15.

La situation est un peu différente dans l'exemple 9.5 où l'on teste l'hypothèse $H_1 = \{\alpha = 0\}$ contre $H_2 = \{\alpha \in \Theta_2\}$ au vu d'un échantillon $X \in \Phi_{\alpha, E}$. On admet que l'ensemble Θ_2 et son adhérence convexe $\bar{\Theta}_2$ ne contiennent pas le point $\alpha = 0$. Si le point $\beta \in \bar{\Theta}_2$ le plus proche de l'origine des coordonnées appartient à Θ_2, il existe un test minimax qui est de la forme

$$X\beta^T > c. \tag{7}$$

Ce test n'est invariant par aucun groupe de transformations. Nous proposons au lecteur de s'assurer que dans ce cas le test du rapport de vraisemblance sera différent de (7) et sera de la forme

$$\rho^2(X, \Theta_2) - \rho^2(X, 0) < c,$$

où

$$\rho(X, \Theta_2) = \inf_{\alpha\in\Theta_2} |X - \alpha|, \rho(X, 0) = |X|.$$

Montrons maintenant que sous certaines conditions le test du rapport de vraisemblance est invariant. Soit G un groupe quelconque de transformations de $\mathscr{X}^n$ laissant invariant le problème de test de H_1 contre H_2, et soit $\bar{G}$ le groupe de transformations $\bar{g}$ sur Θ correspondant.

THÉORÈME 1. *Si $f_\theta(x)$ est telle que*

$$f_\theta(gx) = c(g,x)f_{\bar{g}\theta}(x), \tag{8}$$

le test du rapport de vraisemblance est invariant par G.

Signalons au sujet de la condition (8) qu'elle est toujours remplie si μ est la mesure de Lebesgue et g une transformation préservant cette mesure (une translation, une rotation). Dans ce cas, $c(g,x) \equiv 1$. Pour la contraction, $c(g,x) \equiv$ const.

DÉMONSTRATION du théorème 1. Vu que $\bar{g}\Theta_i = \Theta_i$, $i = 1, 2$, on aura

$$R(gx) = \frac{\sup_{\theta\in\Theta_2} f_\theta(gx)}{\sup_{\theta\in\Theta_1} f_\theta(gx)} = \frac{\sup_{\theta\in\Theta_2} c(g,x)f_{\bar{g}\theta}(x)}{\sup_{\theta\in\Theta_1} c(g,x)f_{\bar{g}\theta}(x)} = \frac{\sup_{\theta\in\bar{g}\Theta_2} f_\theta(x)}{\sup_{\theta\in\bar{g}\Theta_1} f_\theta(x)} = R(x). \blacktriangleleft$$

Les autres propriétés du test du rapport de vraisemblance sont examinées dans les §§ 11, et 13 à 16.

§ 11*. Analyse séquentielle

1. Remarques préliminaires. Jusqu'ici nous avons toujours considéré que la taille n de l'échantillon $X = X_n$ était fixe. Sous cette condition nous avons cherché les tests jouissant de telle ou telle propriété d'optimalité. Par exemple, dans le cas élémentaire où l'on a testé deux hypothèses simples $H_i = \{X \in \mathbf{P}_i\}$, $i = 1, 2$, on a vu qu'il existait un test le plus puissant π de niveau $1 - \epsilon$, de la forme (cf. théorème 2.1)

$$\pi(X) = \begin{cases} 1 & \text{si} \quad f_2(X) > cf_1(X), \\ p & \text{si} \quad f_2(X) = cf_1(X), \\ 0 & \text{si} \quad f_2(X) < cf_1(X), \end{cases}$$

où c et p se déterminent à partir de la condition $\mathbf{E}_1\pi(X) = \epsilon$ et $f_i(x)$ sont les densités des distributions $\mathbf{P}_i$, $i = 1, 2$, par rapport à une mesure μ.

Est-il possible d'améliorer cette procédure statistique ? Certainement pas dans les conditions formulées. Mais si l'on renonce à fixer la taille de l'échantillon, c'est-à-dire si le nombre n d'observations est traité comme

une variable aléatoire dépendant des observations déjà réalisées, alors ces améliorations sont possibles. On entend par là qu'il est possible de réduire le nombre des observations nécessaires à la construction d'un test de paramètres donnés. Cette circonstance est essentielle pour les expériences onéreuses.

Le principe de cette amélioration peut être expliqué sur l'exemple suivant. Supposons que des distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ ne sont pas absolument continues l'une par rapport à l'autre et qu'il existe des ensembles B_1 et B_2 de $\mathfrak{B}_{\mathcal{X}}$ tels que $f_1(x) > 0, f_2(x) = 0$ pour $x \in B_1$, et $f_1(x) = 0, f_2(x) > 0$ pour $x \in B_2$. Il est alors clair que si $\mathrm{x}_1 \in B_1$ (resp. $\mathrm{x}_1 \in B_2$), on peut affirmer indubitablement que l'hypothèse H_1 (resp. H_2) est vraie. Ceci étant, on n'a nul besoin de poursuivre les observations.

Si donc l'on ne procède pas à n observations d'un coup, mais successivement, en tenant compte des résultats des observations précédentes, on peut réduire le nombre n.

L'introduction de la procédure séquentielle est très naturelle du point de vue bayésien. En effet, le test bayésien étudié au § 2 suggère d'accepter l'hypothèse H_2 si la probabilité *a posteriori* $q(2|X)$ de cette hypothèse est $\geqslant 1/2$. Ceci étant, la région critique contiendra entre autres aussi bien des échantillons X pour lesquels $q(2|X)$ est proche de 1 (l'acceptation de H_2 est logique pour de tels X) que des échantillons X pour lesquels $q(2|X)$ est proche de 1/2. Il serait naturel de considérer que ces derniers « ne suffisent pas » pour prendre une décision et impliquent des observations supplémentaires. De plus, de même que dans l'exemple ci-dessus, la probabilité *a posteriori* $q(2|X)$ peut être élevée dès les premières observations et il est alors possible de prendre une décision sans poursuivre les observations (dans l'exemple mentionné, $q(2|X) = 1$ pour $X = \mathrm{x}_1 \in B_2$ pour toute distribution *a priori* $(q(1), q(2))$, $q(2) > 0$).

Nous considérons plus bas une procédure séquentielle de test de deux hypothèses simples donnant lieu à la plus grande réduction possible du nombre d'observations.

2. Test séquentiel bayésien. Commençons par la position bayésienne du problème et désignons par $q(1) = q, q(2) = 1 - q$ les probabilités *a priori* des hypothèses H_1 et H_2. La probabilité *a posteriori* de l'hypothèse H_i après les observations $X = X_n$ sera égale à

$$q(i|X_n) = \frac{q(i)f_i(X_n)}{q(1)f_1(X_n) + q(2)f_2(X_n)} . \tag{1}$$

On réalisera les observations successivement et pour chaque n on calculera les valeurs $q(2|X_n)$, $n = 1, 2, \ldots$ (ou $q(1|X_n)$). Dans le plan (n, y) considérons une *trajectoire aléatoire* des probabilités *a posteriori* (une ligne

polygonale aléatoire) issue du point $y = q(2)$ pour $n = 0$ et prenant aux points $n = 1, 2, \ldots$ les valeurs $y = q(2|X_n)$. Cette trajectoire nous permet de construire le test de H_1 contre H_2 suivant : considérons dans le plan (n, y) deux frontières droites $y = \gamma_i, i = 1, 2$; $0 < \gamma_1 < \gamma_2 < 1$, pour la variable $q(2|X_n)$. On accepte l'hypothèse H_1 ou H_2 selon que la trajectoire $q(2|X_n), n = 0, 1, \ldots$, quitte la première fois la bande (γ_1, γ_2) par la frontière inférieure γ_1 ou supérieure γ_2. Nous verrons plus bas que la $\mathbf{P}_i$-probabilité ($i = 1, 2$) que $q(2|X_n)$ ne quitte jamais la bande (γ_1, γ_2), c'est-à-dire la probabilité de l'événement

$$\{\gamma_1 < q(2|X_n) < \gamma_2, \quad n = 0, 1, \ldots\} \tag{2}$$

est nulle.

Le nombre ν d'observations nécessaire pour accepter l'une des hypothèses (c'est-à-dire pour violer la double inégalité (2)) est visiblement une *variable aléatoire markovienne* (un instant d'arrêt) par rapport à la suite $\mathrm{x}_1, \mathrm{x}_2, \ldots$ pour chacune des distributions $\mathbf{P}_1$ et $\mathbf{P}_2$. De ce point de vue la règle d'acceptation des hypothèses mentionnée est *séquentielle*. Elle s'accorde bien avec les principes qui régissent le comportement de tout un chacun : une décision est prise une fois que les observations permettent de réduire suffisamment l'indétermination qui affecte l'objet étudié.

Le test construit dépend de $q = q(1)$ et du vecteur $\gamma = (\gamma_1, \gamma_2)$. On le notera $\delta_{q,\gamma}$. Montrons qu'il est optimal. A cet effet, introduisons tout d'abord la notion générale de test séquentiel dont les caractéristiques essentielles sont, outre les risques de première et de deuxième espèce, les moyennes $\mathbf{E}_1\nu$ et $\mathbf{E}_2\nu$ du nombre d'observations ν nécessaire à la prise de décision.

Soit donnée sur $(\mathscr{X}^\infty, \mathfrak{B}^\infty_{\mathscr{X}})$ une variable aléatoire $\nu \geqslant 0$ à valeurs entières, markovienne par rapport à la suite $\mathrm{x}_1, \mathrm{x}_2, \ldots$ ($\{\nu \geqslant n\} \in \sigma(\mathrm{x}_1, \ldots \ldots, \mathrm{x}_n) = \mathfrak{B}^n_{\mathscr{X}}$). Appelons $\mathscr{X}^\nu$ l'espace des vecteurs (n, X_n) tels que $\nu(X_\infty) = n, X_n = [X_\infty]_n$. Considérons sur $\mathscr{X}^\nu$ la tribu $\mathfrak{B}^\nu$ engendrée par les événements $\{\nu = n, X_n \in B^n\}$, $B^n \in \mathfrak{B}^n_{\mathscr{X}}$, $n = 0, 1, \ldots$ Il est clair que toute distribution sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$ (ou sur $(\mathscr{X}^\infty, \mathfrak{B}^\infty_{\mathscr{X}})$) induit une distribution sur $(\mathscr{X}^\nu, \mathfrak{B}^\nu)$.

DÉFINITION 1. On appelle *test séquentiel* δ de H_1 contre H_2 un couple (ν, Ω), où $\Omega \in \mathfrak{B}^\nu$ est la région d'acceptation de H_2 (la région critique) et la variable aléatoire ν est supposée être propre par rapport aux distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ ($\mathbf{P}_i(\nu < \infty) = 1, i = 1, 2$).

Dans les cas où l'on aura besoin d'indiquer que ν et Ω se rapportent au test δ on écrira $\nu(\delta)$ et $\Omega(\delta)$.

Il est clair qu'on peut définir de façon équivalente un test séquentiel à l'aide d'une fonction mesurable bivalente sur $\mathscr{X}^\nu$. Il est clair aussi qu'on peut donner un test séquentiel δ en construisant sa région critique (que l'on désignera encore par Ω) dans l'espace $\mathscr{X}^\infty$ tout entier. Mais une telle

application des régions Ω et $\mathscr{X}^{\nu} \setminus \Omega$ d'acceptation des hypothèses H_2 et H_1 dans l'espace $\mathscr{X}^{\infty}$ ne nous fournit pas nécessairement tous les éléments de $\mathscr{X}^{\infty}$: aucune hypothèse n'est acceptée pour ceux d'entre eux qui sont tels que $\nu(X_{\infty}) = \infty$. Mais en vertu de la définition, les $\mathbf{P}_i$-probabilités des ensembles de tels X_{∞} sont nulles.

Tout test non randomisé δ est un cas particulier d'un test séquentiel lorsque $\nu(\delta) \equiv n$ est constant (si $\nu(\delta) \equiv 0$, on prend une décision sans faire d'expériences).

Le test séquentiel δ est, comme tout test ordinaire d'hypothèses simples, caractérisé par les risques $\alpha_i(\delta)$ de i-ième espèce ($i = 1, 2$) :

$$\alpha_i(\delta) = \mathbf{P}_i((\nu, X_{\nu}) \notin \Omega_i),$$

où $\Omega_2 = \Omega, \Omega_1 = \mathscr{X}^{\nu} \setminus \Omega_2$. Par ailleurs, comme déjà signalé, on caractérisera un test séquentiel par les moyennes $\mathbf{E}_i\nu$, $i = 1, 2$. Il est évident que pour un test ordinaire δ construit au vu d'un échantillon X_n on aura $\mathbf{E}_i\nu(\delta) \equiv n$.

Pour tenir compte de l'apparition de ces nouveaux facteurs dans la position du problème (c'est-à-dire des caractéristiques liées à ν) on admettra que la réalisation de chaque observation implique des dépenses chiffrées par la quantité a. Il nous sera commode de caractériser aussi les pertes dues à une fausse décision par des valeurs différentes w_1 et w_2. Plus exactement, on admettra que les pertes de i-ième espèce entraînées par une fausse décision lorsque H_i est vraie sont égales à w_i, $i = 1, 2$.

Avec ces conventions l'espérance mathématique $R(q, \delta)$ des pertes causées par l'utilisation du test δ est égale à

$$R(q, \delta) = q[\alpha_1(\delta)w_1 + a\mathbf{E}_1\nu(\delta)] + (1 - q)[\alpha_2(\delta)w_2 + a\mathbf{E}_2\nu(\delta)]. \quad (3)$$

Cette expression s'appelle *risque bayésien*. Si l'on pose $a = 0$, $w_1 = w_2 = 1$, on obtient une expression pour la probabilité d'erreur du test δ, que nous avons déjà utilisée à maintes reprises dans les §§ 1, 2.

DÉFINITION 2. On appelle *test bayésien* un test séquentiel δ minimisant le risque bayésien (3).

La proposition suivante établit l'optimalité (la bayésienneté) du test $\delta_{q,\gamma}$ construit au début de ce numéro.

THÉORÈME 1. *Pour a, w_1 et w_2 donnés, il existe des γ_1 et γ_2 tels que le test $\delta_{q,\gamma}$ est bayésien.*

DÉMONSTRATION. Désignons par δ_i le test qui conduit à accepter l'hypothèse H_i sans observations, de sorte que $\nu(\delta_i) = 0$, $\alpha_i(\delta_i) = 0$. Voyons d'abord dans quels cas le test δ qui minimise $R(q, \delta)$ est confondu avec δ_1 ou δ_2. Il est évident que

$$R(q, \delta_1) = (1 - q)w_2, \quad R(q, \delta_2) = qw_1.$$

Soit K la classe des tests $\{\delta = \delta(X)\}$ dépendant d'au moins une observation, c'est-à-dire la classe des tests δ tels que $\nu(\delta) \geqslant 1$. Il est évident que $R(q, \delta) \geqslant a$ pour $\delta \in K$. Posons

$$R(q) = \inf_{\delta \in K} R(q, \delta).$$

On a $R(q) < \infty$, puisque le test δ basé sur une seule épreuve ($\nu(\delta) \equiv 1$) appartient à K.

En vertu de la linéarité de $R(q, \delta)$ traitée comme une fonction de q, on a pour tout $p \in]0, 1[$

$$R(pq_1 + (1 - p)q_2) = \inf_{\delta \in K} [pR(q_1, \delta) + (1 - p)R(q_2, \delta)] \geqslant$$

$$\geqslant pR(q_1) + (1 - p)R(q_2).$$

Ceci exprime que $R(q)$ est une fonction concave. Comme $a < R(q) < < \infty$, on en déduit que $R(q)$ est aussi une fonction continue sur $[0, 1]$. Comparons maintenant les risques des tests δ_i et $\delta \in K$ en fonction de q (cf. fig. 5).

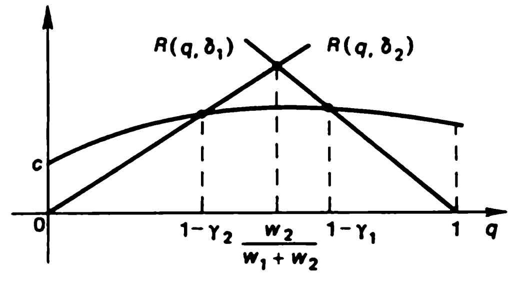


Fig. 5.

De deux choses l'une : ou bien $R(q) \geqslant \min_i R(q, \delta_i)$ pour tous les q $\left(\text{ceci correspond au cas où } R\left(\frac{w_2}{w_1 + w_2}\right) \geqslant \frac{w_1 w_2}{w_1 + w_2}\right)$, ou bien les équations $R(q, \delta_1) = R(q)$ et $R(q, \delta_2) = R(q)$ admettent des solutions que nous désignerons respectivement par $1 - \gamma_1$ et $1 - \gamma_2$, $1 - \gamma_1 > 1 - \gamma_2$. Il est évident que $R(q) < \min_i R(q, \delta_i)$ à l'intérieur de l'intervalle $]1 - \gamma_2, 1 - \gamma_1[$. Pour le premier de ces cas posons

$$1 - \gamma_1 = 1 - \gamma_2 = \frac{w_1}{w_1 + w_2},$$

de sorte que

$$R(1 - \gamma_1, \delta_1) = R(1 - \gamma_1, \delta_2).$$

Des raisonnements précédents et de la figure 5 on déduit la recette optimale suivante. On calcule $1 - \gamma_1$ et $1 - \gamma_2$ à l'aide des valeurs a, w_1 et w_2 qui sont données. Si $q \leqslant 1 - \gamma_2$ ou, ce qui revient au même, $1 - q \geqslant \gamma_2$, c'est le test δ_2 qui fournit le plus petit risque (c'est-à-dire qu'il faut accepter immédiatement l'hypothèse H_2). Si $q \geqslant 1 - \gamma_1$ (ou $1 - q \leqslant \gamma_1$), c'est δ_1 qui donne lieu au plus petit risque (et il faut alors accepter H_1). Et ce n'est que si $1 - \gamma_2 < 1 - \gamma_1$, $q \in]1 - \gamma_2, 1 - \gamma_1[$ (ou $1 - q \in]\gamma_1, \gamma_2[$) qu'il faut utiliser un test de K, c'est-à-dire qu'il faut réaliser une observation.

Raisonnons maintenant par récurrence. Supposons qu'on ait réalisé n observations et que l'on dispose d'un échantillon X_n. Avant la $(n + 1)$-ième observation on est placé devant la même alternative : soit cesser les observations et accepter immédiatement l'une des hypothèses H_i, soit les poursuivre. Le fait que l'on ait déjà subi des pertes an est sans importance, puisqu'on ne peut plus y remédier. Seule la distribution *a priori* fait l'objet de notables changements. Le rôle des probabilités $q(1) = q$ et $q(2) = 1 - q$ incombe maintenant aux probabilités *a posteriori* $q(1|X_n)$ et $q(2|X_n)$. Dans cette nouvelle situation la recette optimale proposée ci-dessus nous commande d'accepter H_2 si $q(2|X_n) \geqslant \gamma_2$ et H_1 si $q(2|X_n) \leqslant \gamma_1$. Si $q(2|X_n) \in]\gamma_1, \gamma_2[$, il faut poursuivre les observations. Mais cette recette n'est autre que celle du test $\delta_{q,\gamma}$. Nous avons donc trouvé des $\gamma_i = \gamma_i(a, w_1, w_2)$ tels que le test $\delta_{q,\gamma}$ minimise le risque $R(q, \delta)$. ◄

Signalons que les nombres $\gamma_i(a, w_1, w_2)$ ne changent pas lorsqu'on multiplie a, w_1 et w_2 par un même nombre : ceci découle de leur définition, puisqu'une telle opération conduit à multiplier tous les risques $R(q, \delta)$ par un même nombre. Donc, γ_i est en fait une fonction de deux variables seulement, par exemple a et w_1 si l'on admet que $w_2 = 1 - w_1$.

Qu'est-ce que le test bayésien $\delta_{q,\gamma}$? Il nous prescrit de ne pas réaliser d'observations dans deux cas : si $\gamma_1 = \gamma_2$ (ce cas se présente lorsque a est grand en regard de w_1 et w_2) ou bien si $q(2) \leqslant \gamma_1$ ou $q(2) \geqslant \gamma_2$. Dans les autres cas, il faut effectuer des expériences jusqu'à la première violation de la double inégalité

$$\gamma_1 < q(2|X_n) < \gamma_2$$

ou, ce qui revient au même, jusqu'à la première violation de

$$\frac{\gamma_1 q(1)}{(1 - \gamma_1)q(2)} < \frac{f_2(X_n)}{f_1(X_n)} < \frac{\gamma_2 q(1)}{(1 - \gamma_2)q(2)} . \tag{4}$$

Ceci étant, on accepte l'hypothèse H_2 si pour la première fois est violée l'inégalité de droite, et l'hypothèse H_1, si c'est celle de gauche. Sous cette forme la partie « variable » du test $\delta_{q,\gamma}$ n'est déjà plus liée à la position bayésienne du problème et nous pouvons désigner par Γ_1 et Γ_2 les bornes

inférieure et supérieure de (4) et considérer un test séquentiel δ_Γ, $\Gamma = (\Gamma_1, \Gamma_2)$, appelé *test séquentiel du rapport de vraisemblance*. Ce test fut introduit par A. Wald.

3. Test séquentiel minimisant le nombre moyen d'observations.

THÉORÈME 2. *Soit* $\Gamma_1 < 1 < \Gamma_2$. *Désignons par* α_1 *et* α_2 *les risques de première et de deuxième espèce du test* δ_Γ. *De tous les tests séquentiels* δ *tels que* $\alpha_1(\delta) \leqslant \alpha_1$, $\alpha_2(\delta) \leqslant \alpha_2$, *le test* δ_Γ *possède les plus petites valeurs* $\mathbf{E}_1\nu(\delta)$ *et* $\mathbf{E}_2\nu(\delta)$.

Ce théorème exprime en particulier que si δ est un test construit au vu d'un échantillon X_n de taille n fixée tel que $\alpha_1(\delta) \leqslant \alpha_1$, $\alpha_2(\delta) \leqslant \alpha_2$, alors

$$\mathbf{E}_i\nu(\delta_\Gamma) \leqslant n, \quad i = 1, 2.$$

DÉMONSTRATION. Le test bayésien $\delta_{q,\gamma}$ envisagé dans le théorème 1 est défini par la collection de nombres (q, a, w_1, w_2). Mais comme déjà signalé, la multiplication de a, w_1 et w_2 par le même nombre ne change pas les bornes γ_i, de sorte que $\delta_{q,\gamma}$ est en fait défini par trois paramètres, par exemple (q, a, w) si l'on convient que $w_1 = w$ et $w_2 = 1 - w$.

Aux termes de cette convention nous avons construit dans le théorème 1 des nombres $\gamma_i = \gamma_i(a, w)$ pour lesquels le test $\delta_{q,\gamma}$ est bayésien. Nous aurons maintenant besoin de la proposition réciproque, savoir que pour γ_1 et γ_2 donnés, il existe des a et w tels que $\gamma_i(a, w) = \gamma_i$, c'est-à-dire des a et w pour lesquels le test $\delta_{q,\gamma}$ est bayésien dans le problème mettant en jeu la collection (q, a, w). Cette proposition revêt un caractère technique et sa démonstration est assez compliquée (cf. [50]). Aussi l'adopterons-nous comme hypothèse *).

Considérons donc le test δ_Γ et pour q donné déterminons γ_i à partir des équations

$$\frac{\gamma_i q}{(1 - \gamma_i)(1 - q)} = \Gamma_i.$$

Pour les valeurs obtenues $\gamma_i = \Gamma_i(1 - q)/(\Gamma_i(1 - q) + q)$ cherchons les a et w pour lesquels le test $\delta_{q,\gamma}$ sera bayésien dans le problème correspondant à la collection (q, a, w). Comme $\Gamma_1 < 1 < \Gamma_2$, il vient $\gamma_1 < 1 - q < \gamma_2$ et $\nu(\delta_{q,\gamma}) \geqslant 1$. Ce qui exprime que $\delta_{q,\gamma} = \delta_\Gamma$.

---

*) Nous ne prouvons pas non plus une autre proposition utile qui dit que pour les $\mathbf{P}_i$-distributions continues de la quantité $f_2(X)/f_1(X)$ et pour des α_1 et α_2 quelconques donnés, il existe des Γ_1 et Γ_2 tels que $\alpha_1(\delta_\Gamma) = \alpha_1$, $\alpha_2(\delta_\Gamma) = \alpha_2$. Cette proposition est voisine des lemmes 6.1 et 7.1, mais sa démonstration est plus compliquée.

Soit maintenant δ un autre test tel que $\alpha_i(\delta) \leqslant \alpha_i$. Puisque le test $\delta_{q,\gamma} = \delta_\Gamma$ minimise le risque bayésien, on a

$$q[\alpha_1 w + a\mathbf{E}_1\nu(\delta_\Gamma)] + (1-q)[\alpha_2(1-w) + a\mathbf{E}_2\nu(\delta_\Gamma)] \leqslant$$
$$\leqslant q[\alpha_1(\delta)w + a\mathbf{E}_1\nu(\delta)] + (1-q)[\alpha_2(\delta)(1-w) + a\mathbf{E}_2\nu(\delta)].$$

D'où il s'ensuit que

$$q\mathbf{E}_1\nu(\delta_\Gamma) + (1-q)\mathbf{E}_2\nu(\delta_\Gamma) \leqslant q\mathbf{E}_1\nu(\delta) + (1-q)\mathbf{E}_2\nu(\delta).$$

Le nombre $q \in]0, 1[$ étant arbitraire, il vient $\mathbf{E}_1\nu(\delta_\Gamma) \leqslant \mathbf{E}_1\nu(\delta)$, $\mathbf{E}_2\nu(\delta_\Gamma) \leqslant \mathbf{E}_2\nu(\delta)$. ◄

Pour la démonstration nous avons appliqué le même procédé de comparaison avec les tests bayésiens que dans les §§ 1, 2, 5.

Considérons quelques *propriétés du test* δ_Γ. Désignons par Ω_i^n les sous-ensembles de $\mathscr{X}^\infty$ définis comme suit ($X_k = [X_\infty]_k$) :

$$\Omega_1^n = \left\{X_\infty : \Gamma_1 < \frac{f_2(X_k)}{f_1(X_k)} < \Gamma_2,\ k = 1, \dots, n-1,\ \frac{f_2(X_n)}{f_1(X_n)} \leqslant \Gamma_1\right\}.$$

L'ensemble Ω_2^n se définit de la même manière, mais il faut remplacer la dernière inégalité par $f_2(X_n)/f_1(X_n) \geqslant \Gamma_2$. Il est évident que les Ω_i^n ne s'intersectent pas, $\Omega_i = \bigcup_{n=1}^{\infty} \Omega_i^n$ est la région d'acceptation de H_i, $\nu(\delta_\Gamma) = n$ dans la région $\{x \in \mathscr{X}^\infty : x \in \Omega_i^n\}$,

$$\alpha_1(\delta_\Gamma) = \sum_{n=1}^{\infty} \mathbf{P}_1(\Omega_2^n) = \sum_{n=1}^{\infty} \int_{\mathscr{X}^n \cap \Omega_i^n} f_1(x)\mu^n(dx) \leqslant$$
$$\leqslant \sum_{n=1}^{\infty} \int_{\mathscr{X}^n \cap \Omega_i^n} f_2(x)\Gamma_2^{-1}\mu^n(dx) = (1 - \alpha_2(\delta_\Gamma))/\Gamma_2. \qquad (5)$$

On établit de façon analogue que

$$\alpha_2(\delta_\Gamma) \leqslant \Gamma_1(1 - \alpha_1(\delta_\Gamma)). \qquad (6)$$

Posons pour simplifier $\alpha_i(\delta_\Gamma) = \alpha_i$. Nous discuterons plus loin du degré de précision des inégalités

$$\Gamma_2 \leqslant \frac{1-\alpha_2}{\alpha_1}, \quad \Gamma_1 \geqslant \frac{\alpha_2}{1-\alpha_1}. \qquad (7)$$

Etablissons maintenant les propriétés du test que l'on obtiendra en se servant des relations (7) pour déterminer Γ_i à l'aide des α_i donnés. Si l'on pose

$$\Gamma_1' = \frac{\alpha_2}{1-\alpha_1}, \quad \Gamma_2' = \frac{1-\alpha_2}{\alpha_1}, \quad \alpha_i' = \alpha_i(\delta_{\Gamma'}),$$

en vertu de (7) on aura pour le test $\delta_{\Gamma'}$ obtenu

$$\frac{\alpha_2}{1-\alpha_1} \geqslant \frac{\alpha_2'}{1-\alpha_1'}, \quad \frac{1-\alpha_2}{\alpha_1} \leqslant \frac{1-\alpha_2'}{\alpha_1'}. \tag{8}$$

D'où

$$\alpha_1' \leqslant \frac{\alpha_1(1-\alpha_2')}{1-\alpha_2} \leqslant \frac{\alpha_1}{1-\alpha_2}, \quad \alpha_2' \leqslant \frac{\alpha_2(1-\alpha_1')}{1-\alpha_1} \leqslant \frac{\alpha_2}{1-\alpha_1}.$$

En réduisant les inégalités (8) au même dénominateur et en les ajoutant, on obtient aussi

$$\alpha_1' + \alpha_2' \leqslant \alpha_1 + \alpha_2.$$

Donc, si les α_i sont petits, le test $\delta_{\Gamma'}$ donnera lieu à des risques α_i' dont la somme sera au plus égale à $\alpha_1 + \alpha_2$ et chacun de ces α_i' ne peut être que légèrement supérieur à α_i et dans des limites que l'on connaît.

EXEMPLE 1. Supposons que x_i suit une distribution binomiale avec une probabilité de succès égale à p. On demande de tester l'hypothèse $H_1 = \{p = p_1\}$ contre $H_2 = \{p = p_2\}$, $p_1 < p_2$. Dans ce cas

$$\frac{f_2(X)}{f_1(X)} = \frac{p_2^{\eta_n}(1-p_2)^{n-\eta_n}}{p_1^{\eta_n}(1-p_1)^{n-\eta_n}} = \left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right)^{\eta_n}\left(\frac{1-p_2}{1-p_1}\right)^n,$$

où η_n est le nombre de succès en n épreuves. Pour $p_1 = 0{,}05$, $p_2 = 0{,}17$, $\alpha_1 = 0{,}05$, $\alpha_2 = 0{,}10$ on obtient *) $\Gamma_1' = 0{,}105$, $\Gamma_2' = 18$, $\alpha_1' = 0{,}031$, $\alpha_2' = 0{,}099$,

$$\mathsf{E}_1\nu(\delta_{\Gamma'}) = 31{,}4, \quad \mathsf{E}_2\nu(\delta_{\Gamma'}) = 30{,}0.$$

A titre de comparaison, si la taille de l'échantillon est fixée et les risques de première et de deuxième espèce égaux à 0,05 et 0,10 respectivement, il faut 57 observations. On voit donc que la procédure séquentielle réduit dans cet exemple presque de deux fois le nombre moyen d'observations.

4. Calcul des paramètres du meilleur test séquentiel. Les relations (7) et (8) permettent d'établir un lien entre la borne Γ et les risques $\alpha_i(\delta_\Gamma)$. Voyons maintenant plus en détail le calcul du test δ_Γ.

*) Les données numériques ont été empruntées à [50].

a) *Formules exactes*. Désignons

$$z_k = \ln \frac{f_2(x_k)}{f_1(x_k)}, \quad k = 1, 2, \dots,$$

$$A_i = \ln \Gamma_i, \quad i = 1, 2.$$

Le test δ_Γ peut alors être mis sous la forme suivante : si $A_1 < 0 < A_2$, on effectue les observations successivement et on somme les valeurs z_k équidistribuées indépendantes jusqu'à ce que la somme $Z_n = \sum_{k=1}^{n} z_k$ atteigne pour la première fois l'une des bornes A_i. Si l'hypothèse H_2 est vraie, la promenade décrite sera dirigée en moyenne vers le haut, puisque

$$\mathsf{E}_2 z_1 = \int \ln \frac{f_2(x)}{f_1(x)} \cdot f_2(x)\mu(dx) = \rho_1(\mathbf{P}_2, \mathbf{P}_1) > 0$$

(cf. lemme 2.6.1). On établit de façon analogue que $\mathsf{E}_1 z_1 = -\rho_1(\mathbf{P}_1, \mathbf{P}_2) < 0$.

Si les bornes A_i s'éloignent de l'origine des coordonnées, ceci correspond (comparer avec (5), (6)) à une baisse des risques de première et de deuxième espèce.

En termes de promenade $\{Z_k\}$ les ensembles Ω_2^n deviennent

$$\Omega_2^n = \{A_1 < Z_k < A_2, k = 1, \dots, n-1, Z_n \geqslant A_2\}.$$

Les ensembles Ω_1^n seront de forme analogue.

Désignons par $\eta(t)$ la variable aléatoire égale à l'instant où la promenade aléatoire $Z_0 = 0$, $Z_1, Z_2, \dots$ traverse pour la première fois le niveau t :

$$\eta(t) = \begin{cases} \min\{k : Z_k \geqslant t\} & \text{pour} \quad t > 0, \\ \min\{k : Z_k \leqslant t\} & \text{pour} \quad t < 0. \end{cases}$$

Ceci est un processus de renouvellement correspondant à la suite $\{Z_k\}$ (cf. [11], chap. 8). Les différences $\chi(A_i) = Z_{\eta(A_i)} - A_i$ caractériseront les dépassements des niveaux A_i dans la promenade $\{Z_k\}$ (cf. [11]).

Pour le risque de première espèce, on peut écrire maintenant

$$\alpha_1(\delta_\Gamma) = \sum_{n=1}^{\infty} \mathbf{P}_1(\Omega_2^n) = \sum_{n=1}^{\infty} \int_{\mathscr{X}^n \cap \Omega_2^n} \frac{f_1(x)}{f_2(x)} f_2(x)\mu^n(dx) =$$

$$= \sum_{n=1}^{\infty} \mathsf{E}_2(e^{-Z_n}; \Omega_2^n) = \Gamma_2^{-1} \mathsf{E}_2(e^{-\chi(A_2)}; \Omega_2), \qquad (9)$$

où $\Omega_2 = \bigcup_{n=1}^{\infty} \Omega_2^n$ est la région d'acceptation de H_2. De façon analogue,

$$\alpha_2(\delta_\Gamma) = \Gamma_1 \mathsf{E}_1(e^{\chi(A_1)}; \Omega_1), \quad \Omega_1 = \bigcup_{n=1}^{\infty} \Omega_1^n. \qquad (10)$$

D'après l'identité de Wald, pour $\mathbf{E}_i\nu$, $i = 1, 2$, $\nu = \nu(\delta_\Gamma)$, on a

$$\mathbf{E}_i(Z_\nu) = \mathbf{E}_i z_1 \mathbf{E}_i \nu, \quad i = 1, 2.$$

Comme $Z_\nu = A_2 + \chi(A_2)$ sur l'ensemble Ω_2 et $Z_\nu = A_1 + \chi(A_1)$ sur Ω_1, il vient

$$\begin{aligned} \mathbf{E}_1\nu &= \frac{1}{\mathbf{E}_1 z_1}\,[\alpha_1 A_2 + \mathbf{E}_1(\chi(A_2)\,;\Omega_2) + (1-\alpha_1)A_1 + \mathbf{E}_1(\chi(A_1)\,;\Omega_1)], \\ \mathbf{E}_2\nu &= \frac{1}{\mathbf{E}_2 z_1}\,[(1-\alpha_2)A_2 + \mathbf{E}_2(\chi(A_2)\,;\Omega_2) + \alpha_2 A_1 + \mathbf{E}_2(\chi(A_1)\,;\Omega_1)]. \end{aligned} \tag{11}$$

Les seconds membres des formules (9), (10) et (11) peuvent être trouvés sous une forme explicite dans de nombreux cas. Ces formules sont d'une grande utilité dans les calculs approchés.

b) *Formules et inégalités approchées (pour de grands A_1 et A_2).* Nous avons déjà signalé que les grands $|A_i|$, $i = 1, 2$, correspondent à de petits risques $\alpha_i(\delta_\Gamma)$. Considérons la valeur

$$\begin{aligned} \alpha_1(\delta_\Gamma) = \mathbf{P}_1(\sup_{k \leq \eta(A_1)} Z_k \geq A_2) = \mathbf{P}_1(\sup_{k \geq 0} Z_k \geq A_2) - \\ - \mathbf{P}_1(\sup_{k \leq \eta(A_1)} Z_k < A_2, \sup_{k > \eta(A_1)} Z_k \geq A_2). \end{aligned} \tag{12}$$

La variable aléatoire $\eta(t)$ étant markovienne, le dernier terme de (12) n'excède pas les valeurs

$$\mathbf{P}_1(\sup_{k > \eta(A_1)} (Z_k - Z_{\eta(A_1)}) \geq A_2 - Z_{\eta(A_1)}) \leq \mathbf{P}_1(\sup_{k \geq 0} Z_k \geq A_2 - A_1).$$

Puisque dans presque tous les cas pratiquement intéressants la probabilité $u(A) = \mathbf{P}_1(\sup_{k \geq 0} Z_k \geq A)$ décroît exponentiellement lorsque A croît (cf. par exemple [26], t. 2. On peut tirer la même conclusion du chap. 10 de [11] où sont exposées les méthodes de calcul de $u(A)$ *)), il vient que pour les grands $|A_i|$ la valeur $u(A_2 - A_1)$ sera d'un ordre de petitesse supérieur à celui de $u(A_2)$. Ceci exprime en vertu de (12) que

$$\alpha_1(\delta_\Gamma) \approx \mathbf{P}_1(\sup_{k \geq 0} Z_k \geq A_2) = u(A_2), \tag{13}$$

de sorte que l'on peut négliger la deuxième borne pour les grands A_1 et A_2 dans (12). On obtient de façon analogue l'approximation

$$\alpha_2(\delta_\Gamma) \approx \mathbf{P}_2(\inf_{k \geq 0} Z_k \leq A_1). \tag{14}$$

Si les $|A_i|$ sont grands et les α_i, petits, les parties principales de (11) nous donnent

$$\mathbf{E}_1\nu \approx \frac{A_1}{\mathbf{E}_1 z_1}, \quad \mathbf{E}_2\nu \approx \frac{A_2}{\mathbf{E}_2 z_1}. \tag{15}$$

On a aussi négligé la deuxième borne en établissant ces formules (qui peuvent être acquises également à l'aide des approximations $\mathbf{E}_i\nu \approx \mathbf{E}_i\eta(A_i) \approx A_i/\mathbf{E}_i z_1$. La dernière relation résulte du théorème de renouvellement ([11])).

---

*) Pour plus de détails voir [9].

La prise en compte des termes suivants rangés par ordre de petitesse dans (11) nous donne

$$\begin{aligned} \mathsf{E}_1\nu &= \frac{1}{\mathsf{E}_1 z_1}(A_1 + \alpha_1(A_2 - A_1) + \mathsf{E}_1\chi_1), \\ \mathsf{E}_2\nu &= \frac{1}{\mathsf{E}_2 z_1}(A_2 - \alpha_2(A_2 - A_1) + \mathsf{E}_2\chi_2), \end{aligned} \tag{16}$$

où α_i sont déterminés par les approximations (12) et (13), et les valeurs $\mathsf{E}_i\chi_i = \lim_{|A_i| \to \infty} \mathsf{E}_i\chi(A_i)$, acquises par les méthodes du chap. 10 de [11].

Considérons maintenant les inégalités (8). Puisque $\chi(A_1) \leqslant 0$ et $\chi(A_2) \geqslant 0$, ces inégalités découlent de (9) et de (10) si l'on remplace $\chi(A_i)$ par 0. Donc, la précision de ces inégalités est définie par l'erreur entraînée par ce changement.

Si les variables aléatoires z_i sont bornées, $b_1 \leqslant z_i \leqslant b_2$, il est alors évident que $\chi(A_2) \leqslant b_2$, $\chi(A_1) \geqslant b_1$, et en plus de (5) et (6) on peut établir les inégalités contraires. Plus exactement

$$\begin{aligned} &\alpha_1(\delta_\Gamma) = \Gamma_2^{-1}\mathsf{E}_2(e^{-\chi(A_2)}; \Omega_2) \geqslant \Gamma_2^{-1}e^{-b_2}(1 - \alpha_2), \\ &\alpha_2(\delta_\Gamma) \geqslant \Gamma_1 e^{b_1}(1 - \alpha_1). \end{aligned} \tag{17}$$

Retournons à l'exemple 1 pour illustrer les relations obtenues. On a

$$Z_n = \eta_n \ln \frac{p_2(1 - p_1)}{p_1(1 - p_2)} + n \ln \frac{1 - p_2}{1 - p_1},$$

où η_n est le nombre de succès en n épreuves. Ceci exprime que pour la $\mathbf{P}_i$-distribution, z_1 prend la valeur $b_2 = \ln(p_2/p_1) \approx 1{,}224$ avec la probabilité p_i et la valeur $b_1 = \ln \dfrac{1 - p_2}{1 - p_1} \approx -0{,}135$ avec la probabilité $1 - p_i$, $i = 1, 2$. D'où il vient

$$\mathsf{E}_1 z_1 = -0{,}067, \quad \mathsf{E}_2 z_1 = 0{,}096, \quad e^{b_2} = 3{,}400, \quad e^{b_1} = 0{,}874.$$

Des deux dernières valeurs, seule la deuxième est proche de 1, de sorte que seule la deuxième des inégalités (17) sera relativement exacte. En utilisant cette inégalité et (7) pour le test $\delta_{\Gamma'}$, on obtient

$$0{,}102 = \frac{\alpha_2'}{1 - \alpha_1'} \leqslant \Gamma_1' \leqslant \frac{\alpha_2'}{(1 - \alpha_1')e^{b_1}} = 0{,}117.$$

Ceci nous donne des bornes assez exactes pour la valeur $\Gamma_1' = 0{,}105$. Dans notre cas

$$A_1' = \ln \Gamma_1' \approx -2{,}254, \quad A_2' = \ln \Gamma_2' \approx 2{,}890.$$

En se servant des formules approchées (15), on obtient pour $\mathsf{E}_i\nu'$, $i = 1, 2$, les valeurs

$$A_1'/\mathsf{E}_1 z_1 = 33{,}639, \quad A_2'/\mathsf{E}_2 z_1 = 30{,}108.$$

Nous voyons que les approximations mêmes les plus grossières, par exemple telles que (15), donnent une idée exacte des valeurs $\mathsf{E}_i\nu'$. Les résultats seront bien plus précis avec les formules (16).

§ 12. Test d'hypothèses multiples dans le cas général

Dans ce paragraphe on n'admettra pas que la distribution de l'échantillon appartient à une famille paramétrique.

Le problème de test de deux hypothèses dans le cas général se pose dans les termes suivants. Soient $\mathscr{P}_1$ et $\mathscr{P}_2$ deux familles de distributions, telles que la distribution $\mathbf{P}$ de X appartienne à $\mathscr{P}_1 \cup \mathscr{P}_2$. On éprouve l'hypothèse $H_1 = \{X \in \mathbf{P}, \mathbf{P} \in \mathscr{P}_1\}$ contre $H_2 = \{X \in \mathbf{P}, \mathbf{P} \in \mathscr{P}_2\}$. Le principe général de construction d'un test (non randomisé *)) $\pi(X) = \delta(X)$ est celui qui a été décrit au § 4 pour le cas paramétrique. Plus exactement, on construit la région critique $\Omega \subset \mathscr{X}^n$ (souvent identifiée à la notion de test) qui est telle que H_2 ou H_1 est acceptée selon que $X \in \Omega$ ou $X \notin \Omega$. Le nombre

$$1 - \epsilon = \inf_{\mathbf{P} \in \mathscr{P}_1} \mathbf{P}(X \notin \Omega)$$

s'appelle *niveau* ou *seuil de signification* du test. La quantité

$$\beta_\pi(\mathbf{P}) = \mathbf{P}(X \in \Omega), \quad \mathbf{P} \in \mathscr{P}_2,$$

est la valeur de la puissance du test π au « point » $\mathbf{P} \in \mathscr{P}_2$.

Comparer les puissances $\beta_\pi(\mathbf{P})$ des tests π lorsque l'ensemble $\mathscr{P}_2$ des contre-hypothèses $\mathbf{P}$ est très riche et construire les tests optimaux dans ces conditions est une tâche très difficile, voire même impossible. Le moins que l'on puisse exiger des tests dans ce cas est que pour tout $\mathbf{P} \in \mathscr{P}_2$ fixe l'on ait

$$\lim_{n \to \infty} \beta_\pi(\mathbf{P}) = 1.$$

DÉFINITION 1. On appelle *convergent* (ou *consistant*) un test π possédant la propriété ci-dessus.

L'essence des tests envisagés, de même que de tous les tests statistiques, correspond au principe fondamental de statistique mathématique évoqué au § 1.4 et au § 2.31. Si ϵ est petit et que l'hypothèse H_1 soit vraie, en utilisant plusieurs fois un test de niveau $1 - \epsilon$ on se trompera (c'est-à-dire on tombera dans la région critique) en moyenne dans 100ϵ % des épreuves seulement. C'est pourquoi nous considérons qu'il est pratiquement impossible de tomber en une seule épreuve dans cette région lorsque H_1 est vraie. De sorte que si l'on se retrouve dans cette région, c'est que l'hypothèse avancée est fausse et l'on rejettera H_1. On dit dans ce cas que les résultats de l'expérience sont en désaccord avec l'hypothèse H_1 du point de vue du test π de niveau $1 - \epsilon$.

---

*) Dans la suite, pour unifier les notations on désignera les tests statistiques par le symbole π, bien que dans ce chapitre ils soient en principe des tests non randomisés.

Les tests de l'hypothèse simple $H_1 = \{X \in \mathbf{P}_1\}$ contre l'hypothèse multiple $H_2 = \{X \in \mathbf{P} \neq \mathbf{P}_1\}$ sont très répandus.

La construction des tests de l'hypothèse simple $H_1 = \{X \in \mathbf{P}_1\}$ repose généralement sur l'« écart » de la distribution empirique $\mathbf{P}_n^*$ par rapport à $\mathbf{P}_1$ au sens d'une certaine « distance » $d(\mathbf{P}, \mathbf{Q})$. Une propriété souhaitable de cette distance est que $d(\mathbf{P}, \mathbf{Q}) = 0$ *seulement* pour $\mathbf{Q} = \mathbf{P}$ et aussi que $d(\mathbf{P}, \mathbf{Q})$ soit continue au « voisinage » du point $\mathbf{Q} = \mathbf{P}$, par exemple pour une métrique uniforme (sinon de petits écarts de $\mathbf{Q}$ par rapport à $\mathbf{P}$ risqueraient de conduire à de grandes valeurs de d). On rappelle que dans le cas paramétrique on a utilisé des considérations analogues pour construire les estimateurs du paramètre inconnu par le minimum de la distance.

Supposons donc que $d(\mathbf{P}, \mathbf{Q})$ est une distance (pas forcément une métrique) sur l'espace des distributions. Supposons que pour $\epsilon > 0$ donné on puisse trouver un $c > 0$ tel que

$$\mathbf{P}_1(d(\mathbf{P}_1, \mathbf{P}_n^*) > c) = \epsilon. \tag{1}$$

On construit un test de la manière suivante :

$$\pi(X) = \begin{cases} 0 & \text{si } d(\mathbf{P}_1, \mathbf{P}_n^*) \leqslant c, \\ 1 & \text{sinon.} \end{cases}$$

Il est évident que π est un test de niveau $1 - \epsilon$.

De même que dans le § 3 on peut introduire la notion de test de *niveau asymptotique* $1 - \epsilon$:

$$\lim_{n \to \infty} \mathbf{P}_1(d(\mathbf{P}_1, \mathbf{P}_n^*) > c) = \epsilon. \tag{2}$$

Les tests décrits sont souvent appelés *tests d'ajustement* (de l'hypothèse $\{X \in \mathbf{P}_1\}$). On peut les construire d'une manière équivalente mais légèrement différente. Soit donnée une fonctionnelle $G(\mathbf{P})$ (ou une suite de fonctionnelles $G_n(\mathbf{P})$) telle que $G(\mathbf{P}) \neq G(\mathbf{P}_1)$ pour $\mathbf{P} \neq \mathbf{P}_1$. On peut alors poser $\pi(X) = 1$ si $|G(\mathbf{P}_n^*) - G(\mathbf{P}_1)| > c$ et $\pi(X) = 0$ sinon, c étant déterminé à partir des mêmes considérations que dans (1) et (2). Il est immédiat de vérifier que cette deuxième approche est équivalente à la première, puisque si l'on connaît G on peut déterminer $d(\mathbf{P}, \mathbf{P}_1) = |G(\mathbf{P}) - G(\mathbf{P}_1)|$ (comparer avec le principe de substitution en théorie de l'estimation), et réciproquement, si la distance $d(\mathbf{P}, \mathbf{P}_1)$ est donnée, on peut construire une fonctionnelle $G(\mathbf{P}) = d(\mathbf{P}, \mathbf{P}_1)$ ($G(\mathbf{P}_1) = 0$) vérifiant les conditions exigées.

Si la fonctionnelle G est de plus telle que $G(\mathbf{P}_n^*) \xrightarrow[p]{} G(\mathbf{P})$ pour $X \in \mathbf{P}$ (c'est toujours le cas si G est une fonctionnelle de type I ou II (cf. § 1.3)), *le test construit sera convergent*. En effet, dans ce cas le nombre $c = c(n)$ réa-

lisant l'égalité (2) doit tendre vers 0 (puisque $\mathbf{P}_1(|G(\mathbf{P}_n^*) - G(\mathbf{P}_1)| > \epsilon) \to 0$ pour tout $\epsilon > 0$), et par conséquent, l'on aura $G(\mathbf{P}_n^*) \underset{P}{\to} G(\mathbf{P})$, $\mathbf{P}(|G(\mathbf{P}_n^*) - G(\mathbf{P}_1)| > c(n)) \to 1$ pour chaque $\mathbf{P}$ fixé $\neq \mathbf{P}_1$.

Considérons maintenant quelques tests d'ajustement bien connus qui sont des réalisations de l'approche décrite ci-dessus.

a) *Test de Kolmogorov*. Soit la statistique (la distance)

$$D(\mathbf{P}_1, \mathbf{P}_n^*) = \sup_t |F_n^*(t) - F(t)|,$$

où $F_n^*(t)$ et $F(t)$ sont des fonctions de répartition associées aux mesures $\mathbf{P}_n^*$ et $\mathbf{P}_1$. Au § 1.8 on a établi que si $F(t)$ est continue et $X \in \mathbf{P}_1$, alors

$$d_K(\mathbf{P}_1, \mathbf{P}_n^*) \equiv \sqrt{n} D(\mathbf{P}_1, \mathbf{P}_n^*) \Rightarrow \sup_{0 \leq t \leq 1} |w^\circ(t)|,$$

où $w^\circ(t)$ est un pont brownien. Ceci entraîne le

THÉORÈME 1 (A. Kolmogorov). *Si $F(t)$ est continue, il existe alors*

$$\lim_{n \to \infty} \mathbf{P}_1(d_K(\mathbf{P}_1, \mathbf{P}_n^*) < x) = K(x) \equiv \mathbf{P}(\sup_{0 \leq t \leq 1} |w^\circ(t)| < x).$$

La fonction $K(x)$ peut être déterminée sous une forme explicite. Elle vaut

$$K(x) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2x^2}.$$

On peut se servir de ce théorème pour construire des tests de niveau asymptotique $1 - \epsilon$. La fonction $K(x)$ est tabulée dans de nombreux ouvrages de statistique mathématique. Pour la valeur ϵ donnée on peut trouver dans une table la constante $c = c_\epsilon$ pour laquelle $K(c) = 1 - \epsilon$. En posant $\pi(X) = 1$ pour $d_k(\mathbf{P}_1, \mathbf{P}_n^*) > c_\epsilon$, on obtient un test d'ajustement de niveau asymptotique $1 - \epsilon$. Il est immédiat de voir qu'il est convergent, puisque la fonctionnelle $G(\mathbf{P}) = \sup_t |F_\mathbf{P}(t) - F(t)|$ (ici $F_\mathbf{P}(t) = \mathbf{P}(]-\infty, t[)$) qui a servi à construire le test de Kolmogorov, est continue par rapport à $F_\mathbf{P}$ pour une métrique uniforme et par suite, est une fonctionnelle de type II (cf. chap. 1), telle que $G(\mathbf{P}_n^*) \underset{\text{p.s.}}{\to} G(\mathbf{P})$ pour $X \in \mathbf{P}$. Reste à se servir des remarques faites ci-dessus sur les conditions de convergence d'un test.

Les résultats du chapitre 1 nous permettent de déterminer le comportement asymptotique de la puissance du test de Kolmogorov par rapport aux contre-hypothèses voisines (cf. § 3). Supposons que $X \in \mathbf{P}$, où la distribution $\mathbf{P}$ admet la fonction de répartition

$$F_\mathbf{P}(x) = F(x) + p(x)n^{-1/2}. \tag{3}$$

On admettra pour simplifier que $p(x)$ est continue et $F(x)$, continue et strictement monotone. La puissance $\beta(\mathbf{P})$ du test de Kolmogorov au « point » $\mathbf{P}$ sera égale à

$$\beta(\mathbf{P}) = \mathbf{P}(d_K(\mathbf{P}_1, \mathbf{P}_n^*) > c) = \mathbf{P}(\sup_t |F(t) - F_n^*(t)|\sqrt{n} > c) =$$

$$= \mathbf{P}(\sup_t |F_{\mathbf{P}}(t) - p(t)n^{-1/2} - F_n^*|\sqrt{n} > c).$$

Le changement de variables $t = F_{\mathbf{P}}^{-1}(u)$, où $F_{\mathbf{P}}^{-1}$ est la fonction réciproque de $F_{\mathbf{P}}$, nous donne

$$\mathbf{P}(\sup_{0 \leqslant u \leqslant 1} |u - p(F_{\mathbf{P}}^{-1}(u))n^{-1/2} - F_n^*(F_{\mathbf{P}}^{-1}(u))|\sqrt{n} > c), \tag{4}$$

où $U_n^*(u) = F_n^*(F_{\mathbf{P}}^{-1}(u))$ est une fonction empirique associée à la distribution $\mathbf{U}_{0,1}$ uniforme sur $[0, 1]$, de sorte que (4) est égale à

$$\mathbf{P}(\sup_{0 \leqslant u \leqslant 1} |u - U_n^*(u) - p(F_{\mathbf{P}}^{-1}(u))n^{-1/2}|\sqrt{n} > c).$$

Par ailleurs, $F_{\mathbf{P}}^{-1}(u) \to F^{-1}(u)$, puisque F est strictement monotone. De là et de la continuité de p, on déduit que

$$\lim_{n \to \infty} \beta(\mathbf{P}) = \mathbf{P}(\sup_{0 \leqslant t \leqslant 1} |w^\circ(t) - a(t)| > c), \text{ où } a(t) = p(F^{-1}(t)). \tag{5}$$

On démontre que cette expression est minimale pour $a(t) \equiv 0$ $(p \equiv 0)$. De ce point de vue, le test de Kolmogorov est asymptotiquement sans biais.

b) *Test de Mises-Smirnov* (*test* ω^2). On conviendra que la distance entre les distributions $\mathbf{P}_1$ et $\mathbf{P}_n^*$ est définie par la statistique

$$\omega_n^2 = d_{\omega^2}(\mathbf{P}_1, \mathbf{P}_n^*) = n \int (F(x) - F_n^*(x))^2 dF(x),$$

qui peut être également utilisée pour construire un test d'ajustement de niveau donné. Au chapitre 1 on a prouvé qu'ici et dans le cas précédent, on a le

THÉORÈME 2. *Il existe une distribution limite telle que*

$$\lim_{n \to \infty} \mathbf{P}_1(\omega_n^2 < x) = \Omega(x) \equiv \mathbf{P}\left(\int_0^1 (w^\circ(t))^2 dt < x\right).$$

La fonction $\Omega(x)$ est de forme très compliquée (cf. [8]). Nous ne l'exhiberons pas ici.

La fonctionnelle

$$G(\mathbf{P}) = \int (F(t) - F_{\mathbf{P}}(t))^2 dF(t)$$

étant une fonctionnelle de type II (§ 1.3), le test ω^2 sera convergent pour les mêmes raisons que dans a).

En appliquant les raisonnements du numéro précédent, on peut établir le comportement asymptotique de la puissance $\beta(\mathbf{P})$ du test ω^2 pour des contre-hypothèses voisines $\mathbf{P}$ de la

forme (3). On trouve de façon analogue que

$$\beta(\mathbf{P}) = \mathbf{P}(\omega_n^2 > c) \to \mathbf{P}\left(\int (w^\circ(t) - a(t))^2 dt > c\right),$$

où $a(t)$ est définie dans (5). La valeur limite obtenue est minimale comme (5) pour $a(t) \equiv 0$, de sorte que le test ω^2 est aussi asymptotiquement sans biais.

Les deux tests considérés, de même que les autres tests d'ajustement de l'hypothèse $H_1 = \{X \in \mathbf{P}_1\}$ construits à l'aide de la distance $d(\mathbf{P}, \mathbf{Q})$, nous permettent d'obtenir immédiatement les ***régions de confiance*** pour la fonction de répartition inconnue $F(x)$ ou pour la distribution inconnue $\mathbf{P}_1$ de l'échantillon X. En effet, la relation (1) (ou (2)) peut être traitée aussi de la manière suivante : la probabilité qu'un c-voisinage du « point » $\mathbf{P}_n^*$ (pour la distance d) recouvre le « point » $\mathbf{P}_1$ est égale à $1 - \epsilon$. (Pour (2) on obtient une version asymptotique de cette assertion.) Ceci exprime (cf. § 8) que le c-voisinage du point $\mathbf{P}_n^*$ est une région de confiance au seuil $1 - \epsilon$ pour la distribution inconnue $\mathbf{P}_1$, $X \in \mathbf{P}_1$. Le test de Kolmogorov par exemple, définit ce voisinage en termes de fonctions de répartition : c'est l'ensemble de toutes les fonctions $F(x)$ telles que

$$\sup_t |F(t) - F_n^*(t)| \leqslant c_\epsilon / \sqrt{n},$$

où c_ϵ se détermine à partir de (1).

Revenons aux tests. Nous avons déjà signalé que l'on ne pouvait faire confiance aux niveaux de signification asymptotiques que pour les grands n. Si la taille de l'échantillon n'est pas élevée, il est nécessaire pour construire les tests (plus exactement, pour trouver $c = c_\epsilon$) de se servir des formules exactes pour la distribution de $d(\mathbf{P}_1, \mathbf{P}_n^*)$. Mais leur acquisition pose en principe de gros problèmes. A cet égard, les tests dits ***non paramétriques*** basés sur des statistiques dont la distribution ne dépend pas de la véritable distribution $\mathbf{P}_1$ (ou ne dépend pas du paramètre θ si $X \in \mathbf{P}_\theta$) jouent un rôle important.

Dans ce cas, les probabilités $\mathbf{P}_1(d(\mathbf{P}_1, \mathbf{P}_n^*) < x)$ ne dépendent pas de $\mathbf{P}_1$, et par conséquent, on peut effectuer les calculs une seule fois, dresser des tables et ensuite les utiliser pour n'importe quelle distribution $\mathbf{P}_1$.

Le test de Kolmogorov et le test ω^2 sont des tests non paramétriques. Ce fait a été établi dans le § 1.6.

Les tests non paramétriques servent aussi à éprouver des ***hypothèses multiples***.

c) ***Test du signe***. Supposons que $F(x)$ est la fonction de répartition de $\mathbf{P}_1$ et que $H_1 = \{F(a) = p\}$, a étant un point donné. Il est évident que H_1 est une hypothèse multiple. L'hypothèse complémentaire est : $H_2 = \{X \in \in \mathbf{P}, F_{\mathbf{P}}(a) \neq p\}$. Dans ce cas il est naturel de se servir de la statistique suivante : désignons par $\nu(X)$ le nombre des observations x_i pour lesquelles le

signe de la différence $x_i - a$ est négatif. Pour région critique Ω, on prendra l'ensemble de tous les échantillons X pour lesquels

$$\nu(X) \notin]c_1, c_2[$$

pour certains $c_1 < c_2$.

Si l'hypothèse H_1 est vraie, on a

$$\mathbf{P}_1(\nu(X) = k) = C_n^k p^k (1 - p)^{n-k}.$$

Si donc l'hypothèse H_1 est vraie, la distribution de $\nu(X)$ est indépendante de $\mathbf{P}_1$ et notre test est non paramétrique. Les nombres c_i doivent être choisis tels que

$$\mathbf{P}(\nu(X) \in]c_1, c_2[) \geqslant 1 - \epsilon$$

(l'égalité peut ne pas être réalisée, car $\nu(X)$ est discrète). L'arbitraire dans le choix de c_i peut être levé par la condition d'absence de biais par rapport aux variations de p. Dans l'ensemble ce problème est équivalent au test de l'hypothèse que la probabilité de succès dans une série d'épreuves de Bernoulli est égale à p. On peut construire de façon identique des tests « unilatéraux » pour éprouver des hypothèses de la forme $F(a) \leqslant p$.

Si l'on généralise le problème posé en considérant l'hypothèse $F(a_i) = = p_i$, $i = 1, \ldots, r$, pour des valeurs a_i et p_i données, on obtient le test du χ^2 qui est étudié en détail dans le § 16.

d) *Test de Moran*. On appelle ainsi le test de l'hypothèse $\{X \in \mathbf{P}_1\}$. Soit $x_{(1)}, \ldots, x_{(n)}$ l'échantillon ordonné associé à X. Supposons que $\mathbf{P}_1$ admet une fonction de répartition F continue et formons la statistique

$$M_n = \sum_{k=0}^{n} [F(x_{(k+1)}) - F(x_{(k)})]^2, \tag{6}$$

où nous conviendrons que $F(x_{(0)}) = 0$, $F(x_{(n+1)}) = 1$. Le test de Moran rejette l'hypothèse $\{X \in \mathbf{P}_1\}$ si $M_n > c$.

Il est évident que ce test n'est pas paramétrique, puisque $F(x_k) \in \mathbf{U}_{0,1}$. Il suffit donc d'envisager le test $M_n > c$ basé sur la statistique

$$M_n = \sum_{k=0}^{n} (x_{(k+1)} - x_{(k)})^2$$

et destiné à vérifier l'hypothèse que la distribution de X est uniforme. L'emploi de la statistique M_n s'impose de lui-même dans ce cas, puisque la quantité $\sum_{i=1}^{n} y_i^2$ atteint son minimum, si $\sum_{i=1}^{n} y_i = 1$, au point $y_1 = \ldots = = y_n = 1/n$.

La proposition suivante peut servir dans le calcul du niveau asymptotique du test de Moran.

THÉORÈME 3. *Si* $X \in \mathbf{P}_1$, *alors*

$$\sqrt{n}(nM_n/2 - 1) \in \Phi_{0,1}.$$

DÉMONSTRATION. Soit $\xi_j \in \mathbf{\Gamma}_{\alpha,1}, j = 1, 2, \ldots$ Alors $\zeta_k = \sum_{j=1}^{k} \xi_j \in \mathbf{\Gamma}_{\alpha,k}$ et en vertu du corollaire 1.6.2, la distribution conjointe des différences

$$x_{(1)}, x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(n-1)}, 1 - x_{(n)}$$

est confondue avec la distribution conjointe de

$$\frac{\xi_1}{\zeta_{n+1}}, \quad \frac{\xi_2}{\zeta_{n+1}}, \quad \ldots, \quad \frac{\xi_{n+1}}{\zeta_{n+1}},$$

de sorte que *)

$$M_n \underset{d}{=} \zeta_{n+1}^{-2} \sum_{j=1}^{n+1} \xi_j^2.$$

La distribution de M_n ne dépend pas de α et l'on peut poser $\alpha = 1$. Alors

$$\mathbf{E}\xi_j^k = \Gamma(k+1) = k!, \mathbf{V}\xi_j = 1, \mathbf{V}\xi_j^2 = 20,$$

$$\rho_n \equiv \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\xi_j - 1) \in \Phi_{0,1},$$

$$\eta_n \equiv \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\xi_j^2 - 2) \in \Phi_{0,20}.$$

On a

$$nM_{n-1} \underset{d}{=} \frac{n\left[2n + \sum_{j=1}^{n} (\xi_j^2 - 2)\right]}{\left(n + \sum_{j=1}^{n} (\xi_j - 1)\right)^2} = \frac{n(2n + \eta_n\sqrt{n})}{(n + \rho_n\sqrt{n})^2} = \frac{2 + \eta_n n^{-1/2}}{(1 + \rho_n n^{-1/2})^2},$$

$$(nM_{n-1} - 2)\sqrt{n} = \frac{\eta_n - 4\rho_n - 2\rho_n^2 n^{-1/2}}{(1 + \rho_n n^{-1/2})^2}, \tag{7}$$

*) Le signe $\underset{d}{=}$ exprime la coïncidence des distributions.

où

$$\eta_n - 4\rho_n = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} (\xi_j' + 2), \quad \xi_j' = \xi_j^2 - 4\xi_j,$$

$$\mathbf{E}\xi_j' = -2, \mathbf{V}\xi_j' = \mathbf{E}(\xi_j^4 - 8\xi_j^3 + 16\xi_j^2) - 4 = 4.$$

Donc, $\eta_n - 4\rho_n \in \Phi_{0,4}$ et en vertu des théorèmes de continuité on déduit de (7) que

$$\sqrt{n}(nM_{n-1}/2 - 1) \in \Phi_{0,1}.$$

Ce qui équivaut à la proposition du théorème. ◁

Montrons maintenant que le test de Moran est convergent. Considérons la statistique (6) pour $X \in \mathbf{P}$, où $\mathbf{P}$ est différente de $\mathbf{P}_1$. Sans nuire à la généralité on peut admettre que l'une des distributions $\mathbf{P}_1$ ou $\mathbf{P}$ ($\mathbf{P}$ pour fixer les idées) est uniforme. Au sujet de F on supposera pour simplifier qu'il existe une densité continue $f(t) = F'(t)$, concentrée sur $[0, 1]$. Alors, pour $X \in \mathbf{U}_{0,1}$ la partie principale de nM_n sera égale à

$$n \sum_{k=0}^{n} [f(x_{(k+1)})(x_{(k+1)} - x_{(k)})]^2 \underset{d}{=} n \sum_{k=1}^{n+1} [f(\zeta_k/\zeta_{n+1})\xi_k/\zeta_{n+1}]^2. \tag{8}$$

La loi forte des grands nombres nous dit que $k^{-1}\zeta_k \underset{\text{p.s.}}{\to} 1$ lorsque $k \to \infty$. Donc, la partie principale (8) sera à son tour égale à

$$\sum_{k=1}^{n} f^2(k/n)\xi_k^2/n. \tag{9}$$

En appliquant de nouveau la loi forte des grands nombres (ou l'inégalité de Tchébychev), on trouve que cette expression converge en probabilité vers

$$2\int_0^1 f^2(t)dt \geqslant 2\left(\int_0^1 f(t)dt\right)^2 = 2.$$

L'inégalité sera stricte si $f(t) \not\equiv 1$. Ceci exprime que pour $X \in \mathbf{P} = \mathbf{U}_{0,1} \neq \mathbf{P}_1$

$$\sqrt{n}(nM_n/2 - 1) \underset{P}{\to} \infty, \quad \text{lorsque } n \to \infty,$$

ce qui entraîne, en vertu du théorème 3, que le test de Moran est convergent pour tout niveau fixé $1 - \epsilon$. ◀

Etant convergent, le test de Moran *ne différencie pas* toutefois *les hypothèses voisines*. Supposons que $X \in \mathbf{P} = \mathbf{U}_{0,1}$,

$$F(t) = t + p(t)n^{-1/2}, \quad t \in [0, 1], \tag{10}$$

$$p(0) = p(1) = 0,$$

et que la fonction $p(t)$ est continûment différentiable. Alors

$$n^{3/2}M_n = n^{3/2}\sum_{k=0}^{n}(x_{(k+1)} - x_{(k)})^2 + 2n\sum_{k=0}^{n}(x_{(k+1)} - x_{(k)})(p(x_{(k+1)}) -$$

$$- p(x_{(k)})) + \sqrt{n}\sum_{k=0}^{n}(p(x_{(k+1)}) - p(x_{(k)}))^2. \qquad (11)$$

La partie principale de la deuxième somme est égale ici à $2n\sum_{k=0}^{n} p'(x_{(k)})(x_{(k+1)} - x_{(k)})^2$, ou pour les mêmes raisons que dans (9)

$$2\sum_{k=1}^{n} p'(k/n)\xi_k^2/n \xrightarrow[P]{} 4\int_0^1 p'(t)dt = 0.$$

Le dernier terme de (11) converge aussi en probabilité vers 0, puisque sa partie principale admet la même distribution que

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}[p'(k/n)]^2\xi_k^2/n,$$

ou que $\frac{2}{\sqrt{n}}\int_0^1 [p'(t)]^2dt \to 0$. Ce qui vient d'être dit exprime que pour la fonction (10) la statistique $n^{3/2}M_n/2 - \sqrt{n}$ aura la même distribution limite $\Phi_{0,1}$ que pour $F(t) = t$. ◄

Signalons que ce fait ne doit pas nous inciter à conclure hâtivement que le test de Moran est mauvais. C'est que s'il ne distingue pas des hypothèses voisines de la forme (10), le test de Moran distingue des hypothèses (voisines dans un certain sens) que les autres tests envisagés dans ce paragraphe ne sont pas en mesure de faire. Nous avons en vue les *hypothèses relatives aux densités*.

Considérons l'hypothèse $H_2 = \{X \in \mathbf{P}\}$, où $\mathbf{P}$ a pour densité

$$f(t) = \begin{cases} 2 & \text{si} \quad 2k\Delta_n \leqslant t < (2k+1)\Delta_n, \\ 0 & \text{si} \quad (2k+1)\Delta_n \leqslant t < (2k+2)\Delta_n, \end{cases} \qquad k = 0, 1, \ldots, N-1,$$

où $\Delta_n = \frac{1}{2N}$, $N = N_n > 0$ est un entier. Pour $\Delta_n = o(n^{-1/2})$, la fonction de répartition $F_{\mathbf{P}}(t)$ de la distribution $\mathbf{P}$ sera telle que

$$\sup_t |F_{\mathbf{P}}(t) - t| = o(n^{-1/2}).$$

Ce qui signifie que l'hypothèse H_2 traitée comme une *hypothèse relative à la fonction de répartition* sera si proche de $H_1 = \{X \in \mathbf{U}_{0,1}\}$ que les tests de Kolmogorov et ω^2 ne pourront pas les discerner asymptotiquement (la valeur limite de la puissance au point $\mathbf{P}$ sera confondue avec le niveau limite du test). Mais les hypothèses H_1 et H_2 traitées comme des *hypothèses relatives aux densités* seront foncièrement différentes, puisque $\sup |f(t) - 1| = 1$. Comme

$x_{(0)} = 0$ et $x_{(n+1)} = 1$, la statistique M_n sera strictement supérieure à $\Delta_n^2 N = \Delta_n/2$ pour $X \in \in \mathbf{P}$. Donc, si $n/N = 2n\Delta_n \to \infty$, $\mathbf{P}$-presque sûrement, on aura

$$nM_n \to \infty.$$

En fixant la région critique $\Omega_2 = \{nM_n > 3\}$, on obtient $\mathbf{P}_1(\Omega_2) \to 0$. Ceci exprime que pour $\Delta_n = o(n^{-1/2})$, $\Delta_n n \to \infty$, le test de Moran discernera les hypothèses H_1 et H_2 avec une probabilité voisine de 1. En d'autres termes, la statistique M_n est sensible aux écarts de la densité, quant au test de Moran on peut le recommander pour éprouver des hypothèses concernant les *densités*. Par ailleurs, nous savons du § 1.10 que les densités empiriques se rapprochent de la densité véritable à une vitesse inférieure à $n^{-1/2}$. Il n'est donc pas étonnant que l'on ne puisse pas discerner des hypothèses qui diffèrent entre elles d'une quantité de l'ordre de $n^{-1/2}$ (cf. 10).

Au sujet du test de Moran et de certains tests considérés précédemment on peut faire une remarque générale. Si l'on compare deux tests de même niveau fixé dont l'un est destiné à traiter un nombre d'alternatives plus grand que l'autre, la puissance du premier pour chaque alternative fixée (rejetée par les deux tests) sera en principe inférieure à celle du second. L'exemple le plus simple illustrant cette circonstance nous est fourni par les tests $|x_1| > \lambda_{\epsilon/2}$ et $x_1 > \lambda_\epsilon$ qui sont destinés à éprouver respectivement les hypothèses $\{\alpha \neq 0\}$ et $\{\alpha > 0\}$ contre $\{\alpha = 0\}$ au vu de l'échantillon $x_1 \in \Phi_{\alpha,1}$. Ici λ_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution $\Phi_{0,1}$. Les puissances au point $\alpha > 0$ seront respectivement égales à

$$1 - \Phi_{0,1}(]-\lambda_{\epsilon/2} - \alpha, \lambda_{\epsilon/2} - \alpha[) < 1 - \Phi(\lambda_\epsilon - \alpha).$$

§ 13. Tests asymptotiquement optimaux. Test du rapport de vraisemblance traité comme un test asymptotiquement bayésien d'une hypothèse simple contre une hypothèse multiple

1. Propriétés asymptotiques du test du rapport de vraisemblance et du test bayésien. Soit à tester l'hypothèse simple $H_1 = \{X \in \mathbf{P}_{\theta_1}\}$ contre l'hypothèse multiple $H_2 = \{X \in \mathbf{P}_\theta \,;\theta \neq \theta_1, \theta \in \Theta\}$. Dans les paragraphes précédents nous avons vu sur des exemples qu'il n'existait pas de test uniformément le plus puissant dans ce cas.

On se place dans l'approche partiellement bayésienne décrite dans les §§ 4 et 9 et qui consiste à admettre que $\theta \in \Theta_2 = \Theta \setminus \{\theta_1\}$ est un paramètre aléatoire de distribution $\mathbf{Q}_2 = \mathbf{Q}$. On peut supposer que $\mathbf{Q}$ est définie sur Θ, $\mathbf{Q}(\{\theta_1\}) = 0$. Dans ce cas la distribution de l'échantillon X sera définie par la densité « moyennisée »

$$f_{\mathbf{Q}}(x) = \int f_t(x)\mathbf{Q}(dt). \tag{1}$$

Donc, si $\mathbf{Q}$ est connue, on peut admettre que l'hypothèse $H_{\mathbf{Q}_2} = H_{\mathbf{Q}}$ selon laquelle X admet une distribution de densité (1), et l'hypothèse H_1 sont des hypothèses simples, et utiliser le lemme de Neyman-Pearson pour construire un test uniformément le plus puissant.

Il se trouve que dans ce cas les tests les plus puissants seront asymptoti-

quement confondus avec le test du rapport de vraisemblance

$$R(X) = \frac{\sup_{\theta \in \Theta} f_\theta(X)}{f_{\theta_1}(X)} = \frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} > c \tag{2}$$

pour « presque toutes » les $\mathbf{Q}$ régulières et par conséquent ne dépendront pas de $\mathbf{Q}$. Ce fait nous permet de considérer que le test trouvé est asymptotiquement optimal au moins dans les cas où l'on peut supposer que $\theta \in \Theta_2$ est aléatoire et que sa distribution $\mathbf{Q}$ est inconnue.

Avant de formuler le théorème correspondant, rappelons quelques résultats utiles et prouvons une proposition auxiliaire dans laquelle le rôle principal sera tenu par les propriétés asymptotiques du rapport de vraisemblance. On étudiera immédiatement le cas d'un paramètre vectoriel ; tous les éléments nécessaires à cette étude figurent dans les §§ 2.28 et 2.29.

Supposons donc que $\theta \in \Theta \subset R^k$, $k \geqslant 1$, et que sont remplies les conditions de régularité (RR) formulées dans le § 2.28. Supposons par ailleurs que $\mathbf{Q}$ admet une densité $q(t)$ par rapport à la mesure de Lebesgue $\lambda(dt) = dt$.

Le lemme de Neyman-Pearson nous dit que le test non randomisé le plus puissant $\pi_{\mathbf{Q}_2} = \pi_{\mathbf{Q}}$ de H_1 contre $H_{\mathbf{Q}}$ sera de la forme suivante : $\pi_{\mathbf{Q}}(X) = 1$ si

$$X \in \Omega(c) = \left\{x : \frac{f_{\mathbf{Q}}(x)}{f_{\theta_1}(x)} > c\right\}, \quad f_{\mathbf{Q}}(x) = \int q(u) f_u(x)\, du, \tag{3}$$

où $c = c_n$ sera choisi ultérieurement en fonction du niveau du test.

Les tests bayésiens de H_1 contre $H_{\mathbf{Q}}$ seront aussi de cette forme.

Les risques de première et de deuxième espèce seront égaux respectivement à

$$\alpha_1(\pi_{\mathbf{Q}}) = \mathbf{P}_{\theta_1}\left(\frac{f_{\mathbf{Q}}(X)}{f_{\theta_1}(X)} > c\right),$$
$$1 - \beta(\pi_{\mathbf{Q}}) = \int q(t) \mathbf{P}_t \left(\frac{f_{\mathbf{Q}}(X)}{f_{\theta_1}(X)} \leqslant c\right) dt, \tag{4}$$

où $\beta(\pi_{\mathbf{Q}}) = \int_{\{f_{\mathbf{Q}}(x) \leqslant c f_{\theta_1}(x)\}} f_{\mathbf{Q}}(x) \mu^n(dx)$ est la puissance du test le plus puissant.

On peut écrire des expressions identiques pour le test du rapport de vraisemblance $\hat{\pi}$ qui conduit à accepter l'hypothèse $H_{\mathbf{Q}}$ lorsque (2) est rem-

plie :

$$\alpha_1(\hat{\pi}) = \mathbf{P}_{\theta_1}\left(\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} > c\right),$$
$$\alpha_2(\hat{\pi}) = \int q(t)\mathbf{P}_t\left(\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} \leqslant c\right)dt = \int\limits_{\{f_{\hat{\theta}^*}(x) \leqslant cf_{\theta_1}(x)\}} f_Q(x)\mu^n(dx). \tag{5}$$

Posons $I = I(\theta_1)$ (la valeur de la matrice d'information de Fisher au point θ_1)

$$\frac{f_Q(X)}{f_{\theta_1}(X)} \equiv \left(\frac{2\pi}{n}\right)^{k/2}\frac{q(\theta_1)}{\sqrt{|I|}}e^{T(X)},$$
$$\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} \equiv e^{\hat{T}(X)}. \tag{6}$$

Les régions critiques des tests π_Q et $\hat{\pi}$ (cf. (3), (2)) peuvent alors être écrites respectivement sous la forme

$$T(X) > c_Q, \quad \hat{T}(X) > \hat{c}. \tag{7}$$

LEMME 1. *Si les conditions* (RR) *du* § 2.28 *sont remplies,* $X \in \mathbf{P}_{\theta_1}$ *et* θ_1 *est un point intérieur de* Θ, *alors*

$$2T(X) = 2\hat{T}(X)(1 + \epsilon_n(X)) \in \mathbf{H}_k,\ \epsilon_n(X) \underset{\text{p.s.}}{\to} 0.$$

DÉMONSTRATION. Ce lemme est la conséquence évidente des théorèmes 2.28.4 et 2.28.5. Il suffit seulement de remarquer que dans les notations du théorème 2.28.4 $\hat{T}(X)$ n'est autre que $Y(u^*)$ (pour $\theta = \theta_1$). ◄

2. Le test du rapport de vraisemblance est asymptotiquement bayésien. Passons à l'énoncé de la proposition principale. On rappelle que lorsqu'on étudie les propriétés asymptotiques des tests, on a en vue non pas un seul mais toute une suite de tests $\pi = \pi_n$, où π_n est un test basé sur l'échantillon X_n. Nous avons eu affaire à une situation analogue en étudiant les propriétés asymptotiques des estimateurs. Ici et ultérieurement — partout où cela sera nécessaire — par test π on comprendra une suite de fonctions $\pi_n(X_n)$ définies pour chaque n et $X_n = [X_\infty]_n$.

DÉFINITION 1. On dit qu'un test π de $H_1 = \{\theta \in \Theta_1\}$ contre $H_2 = \{\theta \in \Theta_2\}$ *appartient à la classe* $\bar{K}_\epsilon$ *des tests de niveau asymptotique* $1 - \epsilon$ si

$$\limsup_{n\to\infty}\sup_{\theta\in\Theta_1}\mathbf{E}_\theta\pi(X) \leqslant \epsilon. \tag{8}$$

Si l'hypothèse H_1 est simple et $\Theta_1 = \{\theta_1\}$, la relation (8) se transforme en l'inégalité

$$\limsup_{n\to\infty}\mathbf{E}_{\theta_1}\pi(X) \leqslant \epsilon.$$

Soit h_ϵ le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à k degrés de liberté ($\mathbf{H}_k(]h_\epsilon, \infty[) = \epsilon$). Du lemme 1 il s'ensuit alors que $\pi_Q \in \tilde{K}_\epsilon$, $\hat{\pi} \in \tilde{K}_\epsilon$ si $c_Q = \hat{c} = h_\epsilon/2$.

DÉFINITION 2. Posons $c_Q = h_\epsilon/2$, de sorte que $\pi_Q \in \tilde{K}_\epsilon$. On dit qu'un test $\pi \in \tilde{K}_\epsilon$ est un *test asymptotiquement bayésien* de $H_1 = \{\theta = \theta_1\}$ contre H_Q si les risques de deuxième espèce relatifs à l'hypothèse H_Q vérifient la relation

$$\limsup_{n\to\infty} \frac{\alpha_2(\pi)}{\alpha_2(\pi_Q)} = \limsup_{n\to\infty} \frac{1-\beta(\pi)}{1-\beta(\pi_Q)} = \limsup_{n\to\infty} \frac{\mathbf{E}_Q(1-\pi(X))}{\mathbf{E}_Q(1-\pi_Q(X))} = 1.$$

Nous avons utilisé dans cette définition le rapport (et non pas la différence) des risques de deuxième espèce, puisque $\alpha_2(\pi_Q) \to 0$ pour $n \to \infty$.

THÉORÈME 1. *Si les conditions (RR) sont remplies et θ_1 est un point intérieur de Θ, le test du rapport de vraisemblance $\hat{\pi}$ (cf. (2) et (7)) appartient à $\tilde{K}_\epsilon$ pour $\hat{c} = h_\epsilon/2$ et est un test asymptotiquement bayésien de H_1 contre H_Q pour toute distribution* $\mathbf{Q}$ *dont la densité $q(t)$ est continue et strictement positive dans Θ. De plus*

$$\alpha_2(\hat{\pi}) \sim \alpha_2(\pi_Q) \sim \frac{q(\theta_1)}{n^{k/2}\sqrt{|I|}} V_k h_\epsilon^{k/2},$$

où $I = I(\theta_1)$ et V_k est le volume de la boule unité de R^k.

DÉMONSTRATION. Nous avons déjà prouvé que $\hat{\pi} \in \tilde{K}_\epsilon$ pour $\hat{c} = h_\epsilon/2$. Considérons maintenant les risques de deuxième espèce. On a en vertu de (4) et (7)

$$\alpha_2(\pi_Q) = \int_{\{T(x)\leq c_Q\}} f_Q(x)\mu^n(dx) = \mathbf{E}_{\theta_1}\left\{\frac{f_Q(X)}{f_{\theta_1}(X)}\,;\, 2T(X) \leq h_\epsilon\right\} =$$

$$= \left(\frac{2\pi}{n}\right)^{k/2} \frac{q(\theta_1)}{\sqrt{|I|}}\, \mathbf{E}_{\theta_1}\{e^{T(X)}\,;\, 2T(X) \leq h_\epsilon\}.$$

Sous le signe de l'espérance mathématique figure une fonction de $2T$ bornée, continue presque partout par rapport à la distribution limite $\mathbf{H}_k$. Donc, pour $n \to \infty$, $\chi_k^2 \in \mathbf{H}_k$,

$$\mathbf{E}_{\theta_1}(e^{T(X)}\,;\, 2T(X) \leq h_\epsilon) \to \mathbf{E}\left\{e^{\frac{1}{2}\chi_k^2}\,;\, \chi_k^2 \leq h_\epsilon\right\} =$$

$$= (2\pi)^{-k/2} \int_{\{|y|^2\leq h_\epsilon\}} e^{\frac{1}{2}|y|^2 - \frac{1}{2}|y|^2} dy_1 \ldots dy_k = (2\pi)^{-k/2} h_\epsilon^{k/2} V_k.$$

Déterminons maintenant le comportement asymptotique de $\alpha_2(\hat{\pi})$. Posons $A_n = \{X : \pi_Q \neq \hat{\pi}\}$. Le lemme 1 nous dit que $\mathbf{P}_{\theta_1}(A_n) \to 0$. Donc, du théorème 2.29.5 il s'ensuit que pour tout N fixé

$$\sup_{|u| \leqslant N} \mathbf{P}_{\theta+u/\sqrt{n}}(A_n) \to 0. \tag{9}$$

Utilisons la représentation (cf. (5))

$$\alpha_2(\hat{\pi}) = \int q(t)\mathbf{P}_t(\hat{T}(X) \leqslant \hat{c})dt =$$

$$= \int\limits_{|t-\theta_1| \leqslant N/\sqrt{n}} + \int\limits_{|t-\theta_1| > N/\sqrt{n}} \leqslant \int q(t)\mathbf{P}_t(T(X) \leqslant \hat{c})dt +$$

$$+ \int\limits_{|t-\theta_1| \leqslant N/\sqrt{n}} q(t)\mathbf{P}_t(A_n)dt + \int\limits_{|t-\theta_1| > N/\sqrt{n}} q(t)\mathbf{P}_t(\hat{T}(X) \leqslant \hat{c})dt.$$

Il vient en vertu de (9)

$$\operatorname*{lim.sup}_{n \to \infty} n^{k/2}\alpha_2(\hat{\pi}) \leqslant \lim_{n \to \infty} n^{k/2}\alpha_2(\pi_Q) +$$

$$+ \max_t q(t) \cdot \operatorname*{lim.sup}_{n \to \infty} \int\limits_{|t-\theta_1| > N/\sqrt{n}} \mathbf{P}_t\left(\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} \leqslant e^{\hat{c}}\right)dt.$$

Mais la probabilité sous le signe d'intégration est au plus égale à

$$\mathbf{P}_t\left(\frac{f_{\theta_1}(X)}{f_t(X)} \geqslant e^{-\hat{c}}\right) \leqslant \exp\{\hat{c}/2 - |t - \theta_1|^2 ng/2\}. \tag{10}$$

Nous nous sommes servis du théorème 2.28.1. Donc, cette intégrale est inférieure à

$$e^{\hat{c}/2} \int\limits_{|u| > N} e^{-|u|^2 g/2}du \to 0$$

pour $N \to \infty$. D'où

$$\operatorname*{lim{\cdot}sup}_{n \to \infty} n^{k/2}\alpha_2(\hat{\pi}) \leqslant \lim_{n \to \infty} n^{k/2}\alpha_2(\pi_Q). \tag{11}$$

Il est évident que cela revient à dire que $\hat{\pi}$ est un test asymptotiquement bayésien.

Reste à établir seulement que $\alpha_2(\hat{\pi}) \sim \alpha_2(\pi_Q)$ ou, ce qui est équivalent en vertu de (11), que

$$\operatorname*{lim{\cdot}inf}_{n \to \infty} n^{k/2}\alpha_2(\hat{\pi}) \geqslant \lim_{n \to \infty} n^{k/2}\alpha_2(\pi_Q). \tag{12}$$

Remarquons à cet effet que le test π_Q que nous avons construit est un test bayésien associé à la probabilité *a priori* q_1 de l'hypothèse H_1, définie à

partir de l'équation (comparer avec (3) et (6))

$$\frac{q_1}{1 - q_1} = \left(\frac{2\pi}{n}\right)^{k/2} \frac{q(\theta_1)}{\sqrt{|I|}} e^{c}.$$

Ceci exprime que le risque de π_Q aura le même comportement asymptotique que

$$\epsilon q_1 + (1 - q_1)\alpha_2(\pi_Q) \sim \epsilon q_1 + \alpha_2(\pi_Q).$$

Si l'on admet que (12) n'est pas vraie, on obtiendra un test $\hat{\pi}$ dont le risque sera inférieur. Ce qui est impossible. ◄

Ces raisonnements montrent que la principale contribution aux risques de deuxième espèce provient des valeurs aléatoires θ qui tombent dans un $n^{-1/2}$-voisinage du point θ_1 (ceci explique l'ordre de petitesse $n^{-k/2}$ de ces probabilités).

En modifiant légèrement les raisonnements de la démonstration du théorème 1 on est conduit à la proposition suivante.

THÉORÈME 2. *Les tests π' et π'' de régions critiques*

$$\begin{aligned} \Omega' &= \{x \in \mathscr{X}^n : n(\hat{\theta}^* - \theta_1)I(\theta_1)(\hat{\theta}^* - \theta_1)^T > h_\epsilon\}, \\ \Omega'' &= \{x \in \mathscr{X}^n : L'(X, \theta_1)I^{-1}(\theta_1)(L'(X, \theta_1))^T > h_\epsilon\} \end{aligned} \quad (13)$$

sont, comme le test $\hat{\pi}$, des tests asymptotiquement bayésiens dans $\tilde{K}_\epsilon$. Cette propriété est préservée si l'on remplace $I(\theta_1)$ par $I(\hat{\theta}^*)$ dans (13).

On obtient les tests (13) si l'on développe

$$\ln \frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} = L(X, \hat{\theta}^*) - L(X, \theta_1)$$

en série au voisinage du point $\hat{\theta}^*$ (cf. théorème 2.28.4). La forme du test $\hat{\pi}$ est dans un sens plus commode, car elle n'est pas liée à la dimension.

DÉMONSTRATION. Nous la laissons au soin du lecteur.

Dans le cas scalaire, la région critique Ω' (lorsque $I(\theta_1)$ est remplacée par $I(\hat{\theta}^*)$) est de la forme

$$\Omega' = \left\{|\hat{\theta}^* - \theta_1| > \left[\frac{h_\epsilon}{n|I(\hat{\theta}^*)|}\right]^{1/2}\right\}, \quad (14)$$

où de toute évidence $h_\epsilon = \lambda^2_{\epsilon/2}, \Phi_{0,1}(]-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}[) = 1 - \epsilon$. On voit que le test π' associé à (14), qui est asymptotiquement équivalent à $\hat{\pi}$, peut être interprété de la manière suivante : $\pi'(X) = 1$ *si θ_1 ne tombe pas dans l'intervalle de confiance au seuil asymptotique* $1 - \epsilon$ *pour le paramètre θ, construit à l'aide de l'estimateur du maximum de vraisemblance $\hat{\theta}^*$.*

Cette interprétation est valable de toute évidence dans le cas vectoriel ;

les régions de confiance seront les ellipsoïdes :

$$(\hat{\theta}^* - \theta)I(\hat{\theta}^*)(\hat{\theta}^* - \theta)^T \leqslant n^{-1}h_\epsilon .$$

On voit donc que les estimateurs du maximum de vraisemblance sont étroitement liés aux tests asymptotiquement bayésiens.

EXEMPLE 1. Supposons que $X \in \mathbf{\Pi}_\lambda$ et soit à tester l'hypothèse $H_1 = \{\lambda = \lambda_1\}$ contre $H_2 = \{\lambda \neq \lambda_1\}$. Dans ce cas, $\hat{\lambda}^* = \bar{x}^*$, $I(\lambda) = \lambda^{-1}$ et un test asymptotiquement bayésien sera de la forme

$$(\bar{x} - \lambda_1)^2 > h_\epsilon \lambda_1 / n,$$

où $\mathbf{H}_1(]h_\epsilon, \infty[) = \epsilon$.

EXEMPLE 2. Supposons que $X \in \mathbf{\Phi}_{\alpha, \sigma^2}$ et soit à tester l'hypothèse $H_1 = \{(\alpha, \sigma^2) = (\alpha_1^2, \sigma_1^2)\}$ contre l'hypothèse complémentaire. On a

$$\hat{\alpha}^* = \bar{x}, \hat{\sigma}^{2*} = S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2, \quad I(\alpha, \sigma^2) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & \sigma^{-4}/2 \end{pmatrix}$$

(cf. § 2.16). Donc, un test asymptotiquement bayésien sera de la forme

$$\frac{(\bar{x} - \alpha_1)^2}{\sigma_1^2} + \frac{(S^2 - \sigma_1^2)^2}{2\sigma_1^4} > \frac{h_\epsilon}{n},$$

où $\mathbf{H}_2(]h_\epsilon, \infty[) = \epsilon$.

3. Le test du rapport de vraisemblance est asymptotiquement sans biais. Fermons ce paragraphe en prouvant que le test du rapport de vraisemblance (2) est asymptotiquement sans biais. Rappelons préalablement qu'un test π de $H_1 = \{\theta \in \Theta_1\}$ contre $H_2 = \{\theta \in \Theta_2\}$ est par définition sans biais si

$$\inf_{\theta \in \Theta_2} \mathbf{E}_\theta \pi - \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi \geqslant 0.$$

DÉFINITION 3. On dit qu'un test π est *asymptotiquement sans biais* si

$$\liminf_{n \to \infty} \left(\inf_{\theta \in \Theta_2} \mathbf{E}_\theta \pi - \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \pi \right) \geqslant 0.$$

THÉORÈME 3. *Le test du rapport de vraisemblance* $\hat{\pi}$ (*cf.* (2), (6) *et* (7)) *de* $H_1 = \{\theta = \theta_1\}$ *contre* $H_2 = \{\theta \neq \theta_1\}$ *est asymptotiquement sans biais.*

DÉMONSTRATION. Vu que dans notre cas $\Theta_1 = \{\theta_1\}$ et $\lim_{n \to \infty} \mathbf{E}_{\theta_1} \hat{\pi} = \epsilon$, il suffit de s'assurer que

$$\liminf_{n \to \infty} \inf_{t \in \Theta} \mathbf{E}_t \hat{\pi} = \liminf_{n \to \infty} \inf_{t \in \Theta} \mathbf{P}_t \left(\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} > e^{\hat{c}} \right) \geqslant \epsilon, \tag{15}$$

où $\hat{c} = h_\epsilon / 2$.

De l'estimation (10) il s'ensuit qu'il existe un $N > 0$ tel que

$$\inf_{|t-\theta_1| \geq N/\sqrt{n}} \mathbf{P}_t \left(\frac{f_{\hat\theta^*}(X)}{f_{\theta_1}(X)} > e^{\hat c} \right) > \epsilon.$$

Reste à prouver que

$$\inf_{|t-\theta_1| \leq N/\sqrt{n}} \mathbf{E}_t \hat\pi \to \epsilon.$$

Mais en vertu des théorèmes 2.28.4 et 2.29.3, pour $X \in \mathbf{P}_t$ on a

$$\hat T(X) \Rightarrow \frac{1}{2} (\xi - u) I (\xi - u)^T, \xi \in \Phi_{0, I^{-1}},$$

$$\mathbf{E}_t(\hat\pi) \to \mathbf{P}_t \left(\frac{1}{2} (\xi - u) I (\xi - u)^T > \hat c = h_\epsilon/2 \right)$$

uniformément en u, $|u| \leq N$, $u = \sqrt{n}(t - \theta_1)$. Le second membre atteint son minimum pour $u = 0$. Ce minimum est égal à $\mathbf{P}(\xi I \xi^T > h_\epsilon) = \epsilon$. ◄

§ 14. Tests asymptotiquement optimaux pour hypothèses multiples voisines

1. Position du problème et définitions. Dans le § 3 nous avons discuté deux approches asymptotiques du problème de test de deux hypothèses simples H_1 et H_2. Si l'on admet que ces hypothèses sont fixes, c'est-à-dire ne changent pas lorsque la taille n de l'échantillon X_n croît, le calcul des risques nous conduit à celui des probabilités des grands écarts, de sorte que l'un au moins des risques tend vers 0. Dans la deuxième approche, les hypothèses H_1 et H_2 sont traitées comme des termes d'une suite d'hypothèses se « rapprochant » l'une de l'autre, la vitesse de rapprochement étant choisie de telle sorte que les risques de première et de deuxième espèce convergent vers leurs propres limites (qui sont différentes de 0 et de 1). Nous avons vu que dans le cas paramétrique les valeurs θ_1 et θ_2 du paramètre qui correspondent aux hypothèses H_1 et H_2, doivent différer d'une quantité de l'ordre de $n^{-1/2}$. L'utilisation de l'une ou de l'autre de ces approches dépend des conditions du problème.

Dans le paragraphe précédent, on a étudié une distribution $\mathbf{Q}$ indépendante de n pour la valeur concurrente de θ et comme il fallait s'y attendre on a trouvé que le risque de deuxième espèce converge vers 0 comme $n^{-k/2}$. Ceci est dû au fait que la principale contribution à ce risque est apportée par les hypothèses voisines pour lesquelles l'écart entre θ et θ_1 est de l'ordre de $n^{-1/2}$ (le volume de la région contenant de tels θ sera justement de l'ordre de $n^{-k/2}$).

Dans ce paragraphe, on considère le test d'*hypothèses multiples voisines* dans le cas où les valeurs alternatives du paramètre se rapprochent lors-

que $n \to \infty$. Il apparaît que dans ce cas le problème de test des hypothèses peut dans un certain sens être ramené à un problème bien plus simple pour une distribution normale.

Formulons le problème en termes plus rigoureux. Soit à éprouver l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ contre l'hypothèse $H_2 = \{\theta \in \Theta_2\}$ au vu d'un échantillon $X \in \mathbf{P}_\theta$. Fixons un point intérieur quelconque θ_1 de Θ et posons

$$\theta = \theta_1 + \gamma n^{-1/2}. \tag{1}$$

Supposons maintenant que les ensembles Θ_i sont de la forme

$$\Theta_i = \theta_1 + \Gamma_i n^{-1/2}, \tag{2}$$

où Γ_i sont indépendants de n. La notation (2) exprime que $\theta \in \Theta_i$ si et seulement si $\gamma \in \Gamma_i$ dans (1). Comme au § 3, les hypothèses $H_i = \{\theta \in \Theta_i\}$ sous la condition (1) seront appelées *voisines* (en fait c'est une suite d'hypothèses qui diffèrent d'un n à l'autre).

Le problème de choix entre les hypothèses voisines H_i au vu de $X \in \mathbf{P}_\theta$ sera appelé *problème A*.

Considérons maintenant un autre problème. Supposons que $Y \in \in \Phi_{\gamma, I^{-1}}$ est un échantillon de taille 1 issu d'une distribution normale $\Phi_{\gamma, I^{-1}}$ de vecteur des moyennes γ et de matrice des moments d'ordre deux $I^{-1} = I^{-1}(\theta_1)$, où $I(\theta_1)$ est la matrice d'information de Fisher au point θ_1 pour le problème A. Désignons par $\mathscr{k}_i$ les hypothèses $\{\gamma \in \Gamma_i\}$. Le problème de test des hypothèses $\mathscr{k}_i$ au vu d'une seule observation $Y \in \Phi_{\gamma, I^{-1}}$ sera appelé *problème B*.

Le fait remarquable qui permet de réaliser la réduction signalée ci-dessus consiste en ce qui suit. Soit $\pi(Y)$ un test optimal dans un sens ou dans l'autre (un test uniformément le plus puissant, bayésien, minimax) de $\mathscr{k}_1$ contre $\mathscr{k}_2$ dans le problème B. Supposons comme toujours que $\hat{\theta}^*$ est un estimateur du maximum de vraisemblance dans le problème A et $\gamma^* = (\hat{\theta}^* - \theta_1)\sqrt{n}$. Dans ces conditions, le test $\pi(\gamma^*)$ de H_1 contre H_2 dans le problème A possédera *asymptotiquement* les mêmes propriétés d'optimalité que le test $\pi(Y)$ dans le problème B.

Donc, pour trouver un test asymptotiquement optimal dans le problème A, nous devons considérer un problème B plus simple encore et trouver (si possible) un test π jouissant de la propriété nécessaire d'optimalité. Si maintenant pour observation Y on prend la valeur γ^* et qu'on la porte dans π, on obtient le test cherché dans le problème A.

On pourrait appeler ce fait *critère limite d'optimalité*. Sa signification est relativement simple. On sait, en effet, d'après les résultats du chapitre 2 que pour $X \in \mathbf{P}_\theta$

$$\sqrt{n}(\hat{\theta}^* - \theta)I^{1/2}(\theta) \in \Phi_{0, E}$$

uniformément en θ. Donc, pour $\theta = \theta_1 + \gamma n^{-1/2}$

$$\sqrt{n}(\hat{\theta}^* - \theta_1) - \gamma \in \Phi_{0, I^{-1}(\theta_1)}$$

ou, ce qui est équivalent,

$$\gamma^* \in \Phi_{\gamma, I^{-1}}.$$

La distribution $\Phi_{\gamma, I^{-1}}$ du problème B n'est par conséquent autre que la distribution limite de γ^*. C'est pourquoi le critère limite d'optimalité est assez naturel ; il réduit le problème de test des hypothèses à un problème « limite ». Ce qui est remarquable dans tout cela c'est que cette réduction ne s'accompagne d'aucune perte d'information sensible sur θ : le test optimal dans le problème B le reste dans le problème A.

Pour formuler ce qui vient d'être dit en termes plus rigoureux, introduisons maintenant les principales notions d'optimalité asymptotiques des tests de choix entre hypothèses voisines dans le problème A.

La classe $\tilde{K}_\epsilon$ des tests π de niveau asymptotique $1 - \epsilon$ a été définie dans le paragraphe précédent (définition 2). Pour $\pi \in \tilde{K}_\epsilon$ on a

$$\lim_{n\to\infty}.\sup.\sup_{\theta\in\Theta_1} \mathbf{E}_\theta \pi(X) \leqslant \epsilon.$$

DÉFINITION 1. On dit qu'un test $\pi_1 \in \tilde{K}_\epsilon$ est *asymptotiquement uniformément le plus puissant dans* $\tilde{K}_\epsilon$ si pour tout $\gamma \in \Gamma_2$ et tout $\pi \in \tilde{K}_\epsilon$, on a

$$\lim_{n\to\infty}.\inf\, (\mathbf{E}_\theta \pi_1(X) - \mathbf{E}_\theta \pi(X)) \geqslant 0,$$

où $\theta = \theta_1 + \gamma n^{-1/2} \in \Theta_2$ pour $\gamma \in \Gamma_2$.

Soient données des distributions $\mathbf{\Pi}_i$ sur Γ_i. Ces distributions induisent sur Θ_i des distributions (concentrées dans un $n^{-1/2}$-voisinage du point θ_1) que nous désignerons par $\mathbf{Q}_i$, $i = 1, 2$. Les hypothèses selon lesquelles θ est un paramètre aléatoire de distribution $\mathbf{Q}_i$ seront comme précédemment désignées par $H_{\mathbf{Q}_i}$.

Appelons $\tilde{K}_\epsilon^{\mathbf{Q}_1}$ la classe des tests π tels que

$$\lim_{n\to\infty}.\sup\, \mathbf{E}_{\mathbf{Q}_1} \pi(X) \leqslant \epsilon,$$

où $\mathbf{E}_{\mathbf{Q}_i}$ représente l'espérance mathématique par rapport à la distribution conjointe de θ et X, $\theta \in \mathbf{Q}_i$, $X \in \mathbf{P}_\theta$. Il est évident que $\tilde{K}_\epsilon \subset \tilde{K}_\epsilon^{\mathbf{Q}}$ pour toute $\mathbf{Q}$.

DÉFINITION 2. On dit qu'un test $\pi_1 \in \tilde{K}_\epsilon^{\mathbf{Q}_1}$ de $H_{\mathbf{Q}_1}$ contre $H_{\mathbf{Q}_2}$ est un *test asymptotiquement bayésien dans* $\tilde{K}_\epsilon^{\mathbf{Q}_1}$ si pour tout autre test $\pi \in \tilde{K}_\epsilon^{\mathbf{Q}_1}$ on a

$$\lim_{n\to\infty}.\inf\, (\mathbf{E}_{\mathbf{Q}_2} \pi_1(X) - \mathbf{E}_{\mathbf{Q}_2} \pi(X)) \geqslant 0. \tag{3}$$

On peut donner une définition équivalente d'un test asymptotiquement bayésien dans laquelle au lieu de (3) on exige que

$$\lim_{n\to\infty} . \inf \left(\mathsf{E}_{Q_2}\pi_1(X) - \mathsf{E}_{Q_2}\pi_{Q_1Q_2}(X)\right) \geqslant 0, \tag{4}$$

où $\pi_{Q_1Q_2}$ est un test bayésien de classe $\tilde{K}_\epsilon^{Q_1}$ de choix entre H_{Q_1} et H_{Q_2} (ou ce qui est équivalent un test le plus puissant de H_{Q_1} contre H_{Q_2} de niveau asymptotique $1 - \epsilon$).

A noter que la définition 2 diffère légèrement de la définition du test asymptotiquement bayésien, donnée dans le paragraphe précédent (cf. définition 13.2 qui fait intervenir le rapport des risques et non pas leur différence). Ces définitions sont équivalentes pour la suite de l'exposé mais la dernière est plus commode.

DÉFINITION 3. On dit qu'un test $\pi_1 \in \tilde{K}_\epsilon$ est un *test asymptotiquement minimax dans* $\tilde{K}_\epsilon$ de H_1 contre H_2 si pour tout autre test $\pi \in \tilde{K}_\epsilon$, on a

$$\lim_{n\to\infty} . \inf \left(\inf_{\theta\in\Theta_2} \mathsf{E}_\theta\pi_1(X) - \inf_{\theta\in\Theta_2} \mathsf{E}_\theta\pi(X)\right) \geqslant 0. \tag{5}$$

Si l'on veut que notre étude soit payante, il faut, comme pour les tests minimax ordinaires (cf. § 9), séparer les ensembles Θ_1 et Θ_2 par une zone intermédiaire, sinon les deux limites inférieures de (5) risqueraient d'être égales à ϵ pour tout test asymptotiquement sans biais π.

Les définitions exhibées montrent que la propriété de telle ou telle optimalité asymptotique se distingue de la propriété ordinaire de cette même optimalité par le fait que la différence correspondante est précédée du signe « lim . inf ».

Outre les tests asymptotiquement bayésiens et minimax des classes $\tilde{K}_\epsilon$ et $\tilde{K}_\epsilon^{Q_1}$, on peut étudier des tests asymptotiquement bayésiens et minimax ordinaires. Soit donnée une distribution $Q = q(1)Q_1 + q(2)Q_2$, $q(1) + q(2) = 1$, sur $\Theta = \Theta_1 \cup \Theta_2$. On dit alors qu'un test π_1 est *asymptotiquement bayésien pour la distribution a priori* Q si pour tout autre test π on a

$$\lim_{n\to\infty} . \inf \left[q(1)\mathsf{E}_{Q_1}\pi_1(X) + q(2)\mathsf{E}_{Q_2}(1 - \pi_1(X)) - \right.$$
$$\left. - q(1)\mathsf{E}_{Q_1}\pi(X) - q(2)\mathsf{E}_{Q_2}(1 - \pi(X))\right] \leqslant 0. \tag{6}$$

La moyenne par rapport à Q de la probabilité d'erreur du test π qui figure dans cette inégalité peut être écrite à l'aide de la probabilité $\alpha(\pi, \theta)$ d'erreur au point θ sous la forme $\mathsf{E}_Q\alpha(\pi, \theta)$, où

$$\alpha(\pi, \theta) = \begin{cases} \mathsf{E}_\theta\pi(X) & \text{si} \quad \theta \in \Theta_1, \\ \mathsf{E}_\theta(1 - \pi(X)) & \text{si} \quad \theta \in \Theta_2. \end{cases}$$

L'inégalité (6) devient alors

$$\lim_{n\to\infty} \inf \mathbf{E}_Q[\alpha(\pi_1(X), \theta) - \alpha(\pi(X), \theta)] \leqslant 0.$$

Un test π_1 sera *asymptotiquement minimax* si

$$\lim_{n\to\infty} \inf [\sup_{\theta\in\Theta} \alpha(\pi_1, \theta) - \sup_{\theta\in\Theta} \alpha(\pi, \theta)] \leqslant 0$$

pour tout autre test π.

L'étude des tests asymptotiquement bayésiens (dans $\tilde{K}_\epsilon^{Q_1}$) et asymptotiquement minimax (dans $\tilde{K}_\epsilon$) est au fond la même que celle des tests asymptotiquement bayésiens et minimax ordinaires. Par exemple, un test bayésien de $\tilde{K}_\epsilon^{Q_1}$ est un test bayésien ordinaire pour un $q(1)$ convenable. Dans ce paragraphe on étudiera les tests des classes $\tilde{K}_\epsilon$ et $\tilde{K}_\epsilon^{Q_1}$; les tests asymptotiquement bayésiens et minimax ordinaires seront examinés dans les chapitres suivants (cf. avant-propos) dans le cadre d'une position plus générale du problème.

2. Propositions fondamentales. Pour alléger au possible l'exposé on introduira une condition qui ne modifie en rien le fond du problème et dont on pourra se dédouaner à tout instant : on dispose à cet effet de tous les résultats nécessaires. Plus exactement, on admettra que les ensembles Γ_i sont bornés, c'est-à-dire qu'il existe un $N > 0$ tel que $\Gamma_i \subset \{\gamma : |\gamma| \leqslant N\}$.

DÉFINITION 4. On dit que des tests π_1 et π_2 de choix entre des hypothèses voisines $H_1 = \{\theta \in \Theta_1\}$ et $H_2 = \{\theta \in \Theta_2\}$ au vu d'un échantillon X sont *asymptotiquement équivalents* si

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_1\cup\Theta_2} |\mathbf{E}_\theta\pi_1(X) - \mathbf{E}_\theta\pi_2(X)| = 0. \tag{7}$$

La condition posée nous permet de remplacer le domaine figurant sous le signe sup dans (7) par le domaine $|\theta - \theta_1| \leqslant N/\sqrt{n}$.

Les tests asymptotiquement équivalents π_1 et π_2 jouissent des propriétés suivantes :

1) *Si* $\pi_1 \in \tilde{K}_\epsilon$ *(resp.* $\tilde{K}_\epsilon^{Q_1}$*), alors* $\pi_2 \in \tilde{K}_\epsilon$ *(resp.* $\tilde{K}_\epsilon^{Q_1}$*).*

2) *Si* π_1 *possède l'une des propriétés d'optimalité asymptotique figurant dans les définitions* 1, 2 *et* 3, *il en sera de même du test* π_2.

La première assertion découle de (7) et de l'inégalité

$$\sup_{\theta\in\Theta_1} \mathbf{E}_\theta\pi_2(X) \leqslant \sup_{\theta\in\Theta_1} \mathbf{E}_\theta\pi_1(X) + \sup_{\theta\in\Theta_1} |\mathbf{E}_\theta(\pi_2 - \pi_1)|.$$

La deuxième s'établit de façon analogue. Si par exemple π_1 est un test asymptotiquement minimax, il en sera de même de π_2 en vertu de (7) et de

l'inégalité

$$\inf_{\theta\in\Theta_2} \mathbf{E}_\theta \pi_2(X) \geqslant \inf_{\theta\in\Theta_2} \mathbf{E}_\theta \pi_1(X) - \sup_{\theta\in\Theta_2} |\mathbf{E}_\theta(\pi_2 - \pi_1)|. \blacktriangleleft$$

Les conditions d'équivalence asymptotique des tests sont établies par le

LEMME 1. *Supposons qu'au voisinage d'un point* θ_1 *sont remplies les conditions* (*RR*), $\pi_i(X) = I_{\{T_n(X)+\epsilon_{ni}(X)>c\}}$, $i = 1, 2$, *où pour* $X \in \mathbf{P}_{\theta_1}$ *on a* $\epsilon_{ni}(X) \underset{P_{\theta_i}}{\to} 0$, $T_n(X) \in \mathbf{G}$, *la distribution* $\mathbf{G}$ *étant continue. Les tests* π_1 *et* π_2 *sont alors asymptotiquement équivalents.*

DÉMONSTRATION. $|\mathbf{E}_t\pi_1(X) - \mathbf{E}_t\pi_2(X)| \leqslant \mathbf{P}_t(A_n)$, où l'événement $A_n = \{\pi_1(X) \neq \pi_2(X)\}$ est tel que $\mathbf{P}_{\theta_1}(A_n) = \mathbf{P}_{\theta_1}(T_n(X) + \epsilon_{n1}(X) > c$, $T_n(X) + \epsilon_{n2}(X) \leqslant c) + \mathbf{P}_n(T_n(X) + \epsilon_{n1}(X) \leqslant c$, $T_n(X) + \epsilon_{n2}(X) >$ $> c) \to 0$ pour $n \to \infty$, puisque la distribution limite de T_n est continue. Donc, $\sup_{|t-\theta_1|\leqslant N/\sqrt{n}} \mathbf{P}_t(A_n) \to 0$ en vertu du théorème 2.29.5. $\blacktriangleleft$

Désignons par $\pi_{\Pi_1\Pi_2}(Y)$ le test bayésien de niveau $1 - \epsilon$ du problème B, destiné à éprouver les hypothèses $\mathscr{K}_{\Pi_i}$ selon lesquelles γ est un paramètre aléatoire de distribution Π_i sur Γ_i, $i = 1, 2$. Ce test est de la forme

$$r(Y) \equiv \frac{\int \exp\left\{-\frac{1}{2}(Y-u)I(Y-u)^T\right\}\Pi_2(du)}{\int \exp\left\{-\frac{1}{2}(Y-u)I(Y-u)^T\right\}\Pi_1(du)} > c, \tag{8}$$

où $c = c_\epsilon$ est déterminé à partir de la condition

$$\int \varphi(\gamma, c)\Pi_1(d\gamma) = \epsilon, \quad \varphi(\gamma, c) = \mathbf{P}(r(Y) > c), \quad Y \in \Phi_{\gamma, I^{-1}}. \tag{9}$$

Ces relations expriment visiblement que $\mathbf{E}_{\Pi_1}\pi_{\Pi_1\Pi_2}(Y) = \epsilon$.

On remarquera que $r(y)$ est une fonction analytique de y. En tant que telle, elle ne peut être constante sur un ensemble de mesure de Lebesgue strictement positive ou de mesure $\Phi_{\gamma, I^{-1}}$ (sinon elle serait partout constante, ce qui n'est possible que pour $\Pi_1 = \Pi_2$). Donc, $\mathbf{P}(r(Y) = c) = 0$ pour tout c et la distribution de $r(Y)$ est continue.

Supposons comme précédemment que $\pi_{Q_1Q_2}(X)$ désigne un test bayésien de niveau asymptotique $1 - \epsilon$ dans le problème A.

THÉORÈME 1. *Si les conditions* (*RR*) *sont remplies au voisinage de* θ_1, *le test* $\pi(X) = \pi_{\Pi_1\Pi_2}(\gamma^*)$, $\gamma^* = (\hat\theta^* - \theta_1)\sqrt{n}$, *est asymptotiquement équivalent au test* $\pi_{Q_1Q_2}$ *et par suite est asymptotiquement bayésien.*

De plus

$$\sup_{|\gamma|\leq N} |\mathsf{E}_{\theta_1+\gamma/\sqrt{n}}\pi(X) - \varphi(\gamma, c)| \to 0 \tag{10}$$

pour $n \to \infty$, *où* $\varphi(\gamma, c) = \mathsf{E}_\gamma \pi_{\Pi_1\Pi_2}(Y)$ *est définie dans* (9).

DÉMONSTRATION. Considérons le test bayésien $\pi_{Q_1Q_2}$ du problème A. Il est de la forme

$$T(X) = \frac{\int f_{\theta_1+u/\sqrt{n}}(X)\Pi_2(du)}{\int f_{\theta_1+u/\sqrt{n}}(X)\Pi_1(du)} > c.$$

Si $X \in \mathbf{P}_{\theta_1}$, le théorème 2.28.5 nous donne

$$T(X) = r(\gamma^*)(1 + \epsilon(X, \theta_1))$$

($\gamma^* = u^*$ pour $\theta = \theta_1$). La distribution de $r(Y)$ est continue, car $\gamma^* \Rightarrow Y \in$ $\in \Phi_{0, I-1}$, ce qui, en vertu du lemme 1, prouve la première proposition du théorème, puisque le test π est de la forme $r(\gamma^*) > c$.

La relation (10) découle de la représentation

$$\mathsf{E}_{\theta_1+\gamma/\sqrt{n}}\pi(X) = \mathsf{E}_{\theta_1+\gamma/\sqrt{n}} I_{\{r(\gamma^*)>c\}} \to \mathbf{P}(r(Y) > c),$$

$Y \in \Phi_{\gamma, I}$ et du théorème 2.29.4. ◄

THÉORÈME 2. *Supposons qu'au voisinage du point* θ_1 *sont remplies les conditions* (*RR*) *et que* $\gamma^* = (\hat{\theta}^* - \theta_1)\sqrt{n}$.

Supposons par ailleurs qu'il existe un test minimax $\pi_1(Y)$ *de niveau* $1 - \epsilon$ *de* $\mathfrak{K}_1$ *contre* $\mathfrak{K}_2$ *dans le problème B et que ce test est bayésien*

$$\pi_1(Y) = \pi_{\Pi_1\Pi_2}(Y) \tag{11}$$

pour des distributions a priori Π_1 *et* Π_2 *vérifiant les conditions*

$$\begin{aligned} \mathsf{E}_{\Pi_1}\pi_1(Y) &= \sup_{\gamma\in\Gamma_1} \mathsf{E}_\gamma \pi(Y), \\ \mathsf{E}_{\Pi_2}\pi_1(Y) &= \inf_{\gamma\in\Gamma_2} \mathsf{E}_\gamma \pi(Y), \qquad Y \in \Phi_{\gamma, I-1} \end{aligned} \tag{12}$$

(*comparer avec les hypothèses du théorème* 9.1). *Alors le test* $\pi(X) =$ $= \pi_{\Pi_1\Pi_2}(\gamma^*)$ *sera asymptotiquement minimax dans la classe* $\tilde{K}_\epsilon$ *des tests de* H_1 *contre* H_2 *dans le problème* A.

DÉMONSTRATION. Le test π_1 étant de niveau $1 - \epsilon$, on a

$$\sup_{\gamma\in\Gamma_1} \mathsf{E}_\gamma \pi_1(Y) = \mathsf{E}_{\Pi_1}\pi(Y) = \epsilon.$$

En vertu de (10) et (12) on en déduit

$$\lim_{n\to\infty} . \sup_{\gamma\in\Gamma_1} \mathsf{E}_{\theta_1+\gamma/\sqrt{n}}\pi_{Q_1Q_2}(X) = \lim_{n\to\infty} \mathsf{E}_{Q_1}\pi_{Q_1Q_2}(X) = \epsilon .$$

Ce qui exprime que $\pi_{Q_1Q_2} \in \tilde{K}_\epsilon$, $\pi_{Q_1Q_2} \in \tilde{K}_\epsilon^{Q_1}$.

Il faut prouver maintenant que pour tout test $\pi^* \in \tilde{K}_\epsilon$

$$\lim_{n\to\infty}.\inf \left(\inf_{\theta\in\Theta_2} \mathsf{E}_\theta \pi(X) - \inf_{\theta\in\Theta_2} \mathsf{E}_\theta\pi^*(X)\right) \geqslant 0.$$

On a

$$\lim_{n\to\infty}\cdot\sup . \inf_{\theta\in\Theta_2} \mathsf{E}_\theta\pi^*(X) \leqslant \lim_{n\to\infty}.\sup \mathsf{E}_\theta\pi^*(X) \leqslant \lim_{n\to\infty}.\sup \mathsf{E}_{Q_2}\pi_{Q_1Q_2}(X). \tag{13}$$

La dernière inégalité est vérifiée, puisque $\pi_{Q_1Q_2}$ est bayésien (c'est-à-dire que $q_1\mathsf{E}_{Q_1}\pi_{Q_1Q_2} + (1 - q_1)\mathsf{E}_{Q_2}(1 - \pi_{Q_1Q_2})$ est minimale pour un q_1 convenable) et que lim sup $\mathsf{E}_{Q_1}\pi^*(X) \leqslant \epsilon$, lim $\mathsf{E}_{Q_1}\pi_{Q_1Q_2} = \epsilon$.

Par ailleurs, d'après (10), (12) et le théorème 1, le dernier membre de (13) est égal à

$$\lim_{n\to\infty} \mathsf{E}_{Q_2}\pi_1(\gamma^*) = \mathsf{E}_{\Pi_2}\pi_{\Pi_1\Pi_2}(Y) =$$

$$= \inf_{\gamma\in\Gamma_2} \mathsf{E}_\gamma\pi_{\Pi_1\Pi_2}(Y) = \lim_{n\to\infty}.\inf_{\gamma\in\Gamma} \mathsf{E}_{\theta_1+\gamma/\sqrt{n}}\pi_{Q_1Q_2}(X). \lhd$$

THÉORÈME 3. *Supposons que dans le problème B il existe un test uniformément le plus puissant* $\pi_1(Y)$ *de niveau* $1 - \epsilon$ *de* $_1$ *contre* $_2$. *Supposons par ailleurs que pour tout* $\gamma_2 \in \Gamma_2$ *il existe une distribution* Π_1 *sur* Γ_1 *telle que*

$$\pi_1(Y) = \pi_{\Pi_1\Pi_2}(Y) \tag{14}$$

est un test bayésien de $_{\Pi_1}$ *contre* $_{\Pi_2}$ *(*Π_2 *est ici concentrée au point* γ_2*). Le test* $\pi(X) = \pi_1(\gamma^*)$ *est alors un test asymptotiquement uniformément le plus puissant (de niveau asymptotique* $1 - \epsilon$*) de* H_1 *contre* H_2 *dans le problème* A.

Signalons que la condition (14) est toujours remplie pour les problèmes des §§ 5, 6 et 7. Ceci résulte de la construction même des tests uniformément les plus puissants effectuée dans ces paragraphes.

DÉMONSTRATION du théorème 3. La relation $\pi_1(\gamma^*) \in \tilde{K}_\epsilon$ découle du théorème 1, puisque

$$\lim_{n\to\infty}.\sup_{\theta\in\Theta_1} \mathsf{E}_\theta\pi_1(\gamma^*) = \sup_{\theta\in\Theta_1}.\lim_{n\to\infty} \mathsf{E}_\theta\pi_1(\gamma^*) = \sup_{\gamma\in\Gamma_1}\varphi(\gamma, c) \leqslant \epsilon.$$

Soit maintenant π^* un autre test de $\tilde{K}_\epsilon$. Alors

$$\lim_{n\to\infty}.\sup \mathsf{E}_{Q_1}\pi^*(X) \leqslant \lim_{n\to\infty}.\sup.\sup_{\theta\in\Theta_1} \mathsf{E}_\theta \pi^*(X) \leqslant \epsilon,$$

et π^* peut par conséquent être traité aussi comme un test de $\tilde{K}_\epsilon^{Q_1}$ de choix entre H_{Q_1} et H_{Q_2}, où Q_1 est induite par la distribution Π_1 (cf. énoncé du théorème) et Q_2 est concentrée au point $\theta_2 = \theta_1 + \gamma_2 n^{-1/2}$. Si $\pi_{Q_1Q_2}$ est un test bayésien de niveau asymptotique $1 - \epsilon$ pour ces distributions, alors

$$\lim_{n\to\infty} \mathsf{E}_{\theta_2}\pi_{Q_1Q_2}(X) \geqslant \lim_{n\to\infty}.\sup \mathsf{E}_{\theta_2}\pi^*(X).$$

Mais le premier membre de cette inégalité est confondu, en vertu du théorème 1, avec la valeur

$$\lim_{n\to\infty} \mathsf{E}_{\theta_2}\pi_{\Pi_1\Pi_2}(\gamma^*) = \lim_{n\to\infty} \mathsf{E}_{\theta_2}\pi_1(\gamma^*). \blacktriangleleft$$

On peut chercher de façon analogue un test asymptotiquement uniformément le plus puissant dans la classe des tests asymptotiquement sans biais.

REMARQUE 1. Si les distributions Π_1 et Π_2 sont concentrées aux points respectifs γ_1 et γ_2, on a

$$r(Y) = \frac{\exp\left\{-\frac{1}{2}(Y-\gamma_2)I(Y-\gamma_2)^T\right\}}{\exp\left\{-\frac{1}{2}(Y-\gamma_1)I(Y-\gamma_1)^T\right\}}.$$

La région critique de $\pi_{\Pi_1\Pi_2}(Y)$ sera donc de la forme

$$YI(\gamma_2-\gamma_1)^T = (YI, (\gamma_2-\gamma_1)) > c.$$

En dimension un, on déduit de là le test asymptotiquement le plus puissant (3.21) étudié au § 3.

REMARQUE 2. Si la distribution Π_1 est concentrée en $u = 0$ et la distribution Π_2 est uniforme sur la boule $|u| \leqslant N$, le dénominateur de la fonction $r(Y)$ sera égal à $\exp\left\{-\frac{1}{2}YIY^T\right\}$ et le numérateur sera, pour les grands N et $|\gamma| < N - \sqrt{N}$, proche de $\sqrt{|I|}(2\pi)^{k/2}$. Pour de telles Π_1 et Π_2 la région critique de $\pi_{\Pi_1\Pi_2}$ sera donc d'une forme voisine de l'extérieur de l'ellipsoïde

$$YIY^T > c,$$

et la région critique du test asymptotiquement bayésien $\pi_{\Pi_1\Pi_2}(\gamma^*)$, de l'extérieur de l'ellipsoïde

$$\gamma^* I\gamma^{*T} > c.$$

On reconnaît ici la forme asymptotique du test du rapport de vraisemblance étudié dans le paragraphe précédent (comparer avec le théorème 13.2).

REMARQUE 3. Les théorèmes 2 et 3 contiennent des conditions stipulant que le test minimax (théorème 2) et le test uniformément le plus puissant (théorème 3) du problème B sont bayésiens pour certaines distributions $\mathbf{\Pi}_i$ sur Γ_i. Nous verrons dans les chapitres suivants que ces conditions sont superflues : la classe des tests bayésiens contient tous les tests « inaméliorables », y compris les tests uniformément les plus puissants et les tests minimax.

§ 15. Propriétés d'optimalité asymptotique du test du rapport de vraisemblance découlant du critère limite d'optimalité

Dans ce paragraphe nous étudierons quelques conséquences des résultats du § 14 relatifs au test du rapport de vraisemblance. Nous établirons en particulier que le test du rapport de vraisemblance est asymptotiquement uniformément le plus puissant et minimax pour certains problèmes importants de décision entre hypothèses voisines.

Dans la suite on admettra que les conditions (RR) sont remplies au voisinage du point θ_1. Pour simplifier les raisonnements on supposera au besoin comme dans le paragraphe précédent que les ensembles Γ_i sont bornés.

1. Test asymptotiquement uniformément le plus puissant pour hypothèses voisines avec des contre-hypothèses unilatérales. Supposons que le paramètre θ est scalaire et considérons le test de choix entre l'hypothèse unilatérale $H_1 = \{\theta \leqslant \theta_1 + \gamma_1 n^{-1/2}\}$ et son alternative $H_2 = \{\theta > \theta_2 = = \theta_1 + \gamma_2 n^{-1/2}\}, \gamma_1 \leqslant \gamma_2$.

THÉORÈME 1. *Le test du rapport de vraisemblance $\hat{\pi}(X)$ de région critique*

$$R(X) = \frac{\sup_{\theta\in\Theta_2} f_\theta(X)}{\sup_{\theta\in\Theta_1} f_\theta(X)} > c \tag{1}$$

avec $\Theta_1 = \{\theta : \theta \leqslant \theta_1 + \gamma_1 n^{-1/2}\}, \Theta_2 = \{\theta : \theta \geqslant \theta_1 + \gamma_2 n^{-1/2}\}$ et c convenablement choisi, est asymptotiquement équivalent au test

$$\gamma^* = (\hat{\theta}^* - \theta_1)\sqrt{n} > c_\epsilon = \lambda_\epsilon I^{-1/2} + \gamma_1, \quad \Phi_{0,1}(\lambda_\epsilon) = 1 - \epsilon \tag{2}$$

et est un test asymptotiquement uniformément le plus puissant de niveau asymptotique $1 - \epsilon$ de l'hypothèse $H_1 = \{\theta \leqslant \theta_1 + \gamma_1 n^{-1/2}\}$ contre l'hypothèse $H_2 = \{\theta > \theta_1 + \gamma_2 n^{-1/2}\}$. Dans les formules (2), *le symbole I désigne la quantité d'information de Fischer $I(\theta_1)$ au point θ_1 pour la famille f_θ.*

DÉMONSTRATION. Du § 5 il s'ensuit que pour l'échantillon $Y \in \Phi_{\gamma, I^{-1}}$ de taille un et de variance I^{-1} connue, il existe un test uniformément le plus puissant de l'hypothèse $\mathscr{k}_1 = \{\gamma \leqslant \gamma_1\}$ contre $\mathscr{k}_2 = \{\gamma > \gamma_2\}$, qui est de la forme $Y > c_\epsilon$, où c_ϵ est défini dans (2). Il est évident que le test bayésien associé aux distributions dégénérées concentrées aux points γ_1 et γ_2 (ou aux points γ_1 et $\gamma > \gamma_1$ si $\gamma_1 = \gamma_2$) sera aussi de la même forme. Le théorème 14.3 nous dit alors qu'il existe un test asymptotiquement uniformément le plus puissant de niveau asymptotique $1 - \epsilon$ de H_1 contre H_2 et qu'il est de la forme (2).

Reste à prouver que les tests (1) et (2) sont asymptotiquement équivalents. En posant $Z_1(t) = \dfrac{f_{\theta_1 + t}(X)}{f_{\theta_1}(X)}$, on trouve en vertu du théorème 2.28.4 que pour $X \in \mathbf{P}_{\theta_1}$

$$R(X) = \frac{\sup\limits_{u > \gamma_2} Z_1(un^{-1/2})}{\sup\limits_{u \leqslant \gamma_1} Z_1(un^{-1/2})} =$$

$$= \frac{\sup\limits_{u > \gamma_2} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I + \epsilon_n^{(2)}(X)\right\}}{\sup\limits_{u \leqslant \gamma_1} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I + \epsilon_n^{(1)}(X)\right\}} = T_n(X) + \epsilon_n^{(3)}(X),$$

où $\epsilon_n^{(i)}(X) \underset{P_{\theta_1}}{\rightarrow} 0,\ i = 1, 2, 3,$

$$T_n(X) = r(\gamma^*) \equiv \frac{\sup\limits_{u > \gamma_2} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I\right\}}{\sup\limits_{u \leqslant \gamma_1} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I\right\}} =$$

$$= \begin{cases} \exp\left\{-\frac{1}{2}(\gamma^* - \gamma_2)^2 I\right\} & \text{pour} \quad \gamma^* \leqslant \gamma_1, \\ \exp\left\{-\frac{1}{2}(\gamma^* - \gamma_2)^2 I + \frac{1}{2}(\gamma^* - \gamma_1)^2 I\right\} & \text{pour} \quad \gamma_1 < \gamma^* < \gamma_2, \\ \exp\left\{\frac{1}{2}(\gamma^* - \gamma_1)^2 I\right\} & \text{pour} \quad \gamma^* \geqslant \gamma_2. \end{cases}$$

On reconnaît ici une fonction continue monotone, strictement croissante de γ^*. Donc, l'inégalité $T_n(X) > c$ est équivalente à $\gamma^* > c'$ pour un certain c'. La distribution de $r(Y)$ est absolument continue, puisque $\gamma^* \Rightarrow$

$\Rightarrow Y \in \Phi_{0, l-1}$. Les conditions du lemme 14.1 sont remplies pour les tests (1) et (2). ◄

2. Test asymptotiquement uniformément le plus puissant pour alternatives bilatérales. Supposons comme toujours que le paramètre θ est scalaire et que le problème A consiste à tester l'hypothèse $H_1 = \{(\theta - \theta_1)\sqrt{n} \notin]\gamma_1, \gamma_2[\}$ contre $H_2 = \{(\theta - \theta_1)\sqrt{n} \in]\gamma_1, \gamma_2[\}, \gamma_2 > \gamma_1$. Posons

$$\bar{\gamma} = \frac{\gamma_1 + \gamma_2}{2}, \quad \Delta = \frac{\gamma_2 - \gamma_1}{2}.$$

THÉORÈME 2. *Le test du rapport de vraisemblance $\hat{\pi}(X)$ défini dans* (1) *pour c convenablement choisi et* $\Theta_1 = \{\theta : (\theta - \theta_1)\sqrt{n} \notin]\gamma_1, \gamma_2[\}$ *et* $\Theta_2 = \{\theta : (\theta - \theta_1)\sqrt{n} \in]\gamma_1, \gamma_2[\}$, *et le test*

$$|\gamma^* - \bar{\gamma}| = |(\hat{\theta}^* - \theta_1)\sqrt{n} - \bar{\gamma}| < c_\epsilon, \tag{3}$$

où c_ϵ est déterminé à partir de l'équation $\Phi_{0, l-1}(]-c - \Delta, c - \Delta[) = \epsilon$, *sont des tests asymptotiquement uniformément les plus puissants de niveau asymptotique* $1 - \epsilon$ *de* $H_1 = \{(\theta - \theta_1)\sqrt{n} \notin]\gamma_1, \gamma_2[\}$ *contre* $H_2 = \{(\theta - \theta_1)\sqrt{n} \in]\gamma_1, \gamma_2[\}$.

DÉMONSTRATION. Elle est calquée sur celle du théorème précédent. Du § 5 il s'ensuit que pour le problème B de décision entre l'hypothèse $\mathfrak{k}_1 = \{\gamma \notin]\gamma_1, \gamma_2[\}$ et son alternative $\mathfrak{k}_2 = \{\gamma \in]\gamma_1, \gamma_2[\}$ au vu d'une observation $Y \in \Phi_{\gamma, l-1}$, il existe un test uniformément le plus puissant de la forme $c' < Y < c''$, où c' et c'' sont choisis de telle sorte que

$$\Phi_{\gamma_1, l-1}(]c', c''[) = \Phi_{\gamma_2, l-1}(]c', c''[) = \epsilon.$$

Il est immédiat de voir que ces relations sont remplies si l'on pose $c' = \bar{\gamma} - c_\epsilon$ et $c'' = \bar{\gamma} + c_\epsilon$, puisque

$$\Phi_{\gamma_1, l-1}(]\bar{\gamma} - c_\epsilon, \bar{\gamma} + c_\epsilon[) = \Phi_{0, l-1}(]-c_\epsilon + \Delta, c_\epsilon + \Delta[) = \epsilon,$$
$$\Phi_{\gamma_2, l-1}(]\bar{\gamma} - c_\epsilon, \bar{\gamma} + c_\epsilon[) = \Phi_{0, l-1}(]-c_\epsilon - \Delta, c_\epsilon - \Delta[) = \epsilon.$$

Nous avons vu par ailleurs au § 5 que pour tout $\gamma_0 \in]\gamma_1, \gamma_2[$ il existe un $q \in]0, 1[$ tel que le test bayésien $\pi_{\Pi_1\Pi_2}$ de l'hypothèse $\mathfrak{k}_{\Pi_1}$ associée à la distribution $\Pi_1 : \Pi_1(\{\gamma_1\}) = q, \Pi_1(\{\gamma_2\}) = 1 - q$, contre l'hypothèse $\mathfrak{k}_{\Pi_2} = \{\gamma = \gamma_0\}$ sera de la forme

$$c' < Y < c''.$$

Ceci exprime que les conditions du théorème 14.3 seront remplies et le test (3) sera un test asymptotiquement uniformément le plus puissant de H_1 contre H_2.

Considérons maintenant le test du rapport de vraisemblance (1) pour les régions Θ_i définies dans le théorème et montrons qu'il est asymptotiquement équivalent à (3). Comme dans la démonstration du théorème 1, pour $X \in \mathbf{P}_{\theta_1}$ le théorème 2.28.4 nous donne

$$\frac{\sup\limits_{u \in]\gamma_1, \gamma_2[} Z_1(un^{-1/2})}{\sup\limits_{u \notin]\gamma_1, \gamma_2[} Z_1(un^{-1/2})} =$$

$$= \frac{\sup\limits_{u \in]\gamma_1, \gamma_2[} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I + \epsilon_n^{(1)}(X)\right\}}{\sup\limits_{u \notin]\gamma_1, \gamma_2[} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I + \epsilon_n^{(2)}(X)\right\}} = T_n(X) + \epsilon_n^{(3)}(X),$$

où $\epsilon_n^{(i)}(X) \underset{\mathbf{P}_{\theta_1}}{\to} 0,\ i = 1, 2, 3,$

$$T(X) = r(\gamma^*) = \frac{\sup\limits_{u \in]\gamma_1, \gamma_2[} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I\right\}}{\sup\limits_{u \notin]\gamma_1, \gamma_2[} \exp\left\{-\frac{1}{2}(\gamma^* - u)^2 I\right\}} =$$

$$= \begin{cases} \exp\left\{-\frac{1}{2}(\gamma^* - \gamma_1)^2 I\right\} & \text{si} \quad \gamma^* \leqslant \gamma_1, \\ \exp\left\{\frac{1}{2}(\gamma^* - \gamma_1)^2 I\right\} & \text{si} \quad \gamma_1 < \gamma^* \leqslant \bar{\gamma}, \\ \exp\left\{\frac{1}{2}(\gamma^* - \gamma_2)^2 I\right\} & \text{si} \quad \bar{\gamma} < \gamma^* \leqslant \gamma_2, \\ \exp\left\{-\frac{1}{2}(\gamma^* - \gamma_2)^2 I\right\} & \text{si} \quad \gamma_2 < \gamma^*. \end{cases}$$

On voit sur ces égalités que $r(\gamma^*)$ est une fonction de $|\gamma^* - \bar{\gamma}|$ continue monotone strictement décroissante. Donc l'inégalité $r(\gamma^*) > c$ est équivalente à l'inégalité $|\gamma^* - \bar{\gamma}| \leqslant c'$. Les conditions du lemme 14.1 sont remplies puisque $\gamma^* \Rightarrow Y \in \Phi_{0, I^{-1}}$. ◄

3. Test asymptotiquement minimax pour hypothèses voisines relatives à un paramètre vectoriel. Considérons maintenant un paramètre vectoriel θ. Dans ce cas, il n'existe pas en général de test asymptotiquement uniformément le plus puissant entre $H_1 = \{\theta \in \Theta_1\}$ et $H_2 = \{\theta \in \Theta_2\}$, par contre il est possible de construire des tests asymptotiquement minimax.

Faisons d'abord une remarque générale qui facilitera les raisonnements ultérieurs, savoir que l'on peut toujours « reparamétriser » le problème de

test envisagé (c'est-à-dire introduire un nouveau paramètre) de telle sorte que la matrice d'information $I = I(\theta_1)$ soit une matrice unité au point θ_1. Il suffit pour cela d'effectuer une transformation linéaire (cf. § 2.1) et d'introduire le nouveau paramètre β à l'aide de l'égalité

$$\theta = \beta I^{-1/2}.$$

La matrice d'information de Fisher $J(\beta)$ pour la famille paramétrique $\mathbf{P}_{\beta I^{-1/2}}$ sera alors égale au point $\beta_1 = \theta_1 I^{1/2}$ à

$$J(\beta_1) = I^{-1/2} I I^{-1/2} = E.$$

Dans ce numéro on se servira parfois du paramètre β par simple raison de commodité. On pourra toujours revenir au paramètre primitif à l'aide de la transformation réciproque.

Soit donc $I = I(\theta_1) = E$. Nous étudions le problème A de test au vu d'un échantillon $X \in \mathbf{P}_\theta$ de

$$H_1 = \{|\theta - \theta_1| \leqslant an^{-1/2}\} \text{ contre } H_2 = \{|\theta - \theta_1| \geqslant bn^{-1/2}\}, a < b. \quad (4)$$

THÉORÈME 3. *Le test du rapport de vraisemblance* $\hat{\pi}$ *défini dans* (1) *pour c convenablement choisi,* $\Theta_1 = \{\theta : |\theta - \theta_1| \leqslant an^{-1/2}\}$ *et* $\Theta_2 = \{\theta : |\theta - \theta_1| \geqslant bn^{-1/2}\}$, *est asymptotiquement équivalent pour tous* $0 \leqslant a < b < \infty$ *aux tests*

$$\frac{f_{\hat{\theta}^*}(X)}{f_{\theta_1}(X)} > c, \quad (5)$$

$$|\gamma^*| \equiv |(\hat{\theta}^* - \theta_1)\sqrt{n}| > c_\epsilon, \quad (6)$$

où c_ϵ^2 *est la solution de l'équation en c*

$$p_c(a) \equiv \mathbf{P}((\xi_1 + a)^2 + \xi_2^2 + \ldots + \xi_k^2 > c^2) = \epsilon, \quad (7)$$

et est un test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$ *entre les hypothèses* H_1 *et* H_2 *définies dans* (4). *Les variables aléatoires* ξ_i *dans* (7) *sont indépendantes et* $\xi_i \in \Phi_{0,1}$. *La puissance limite garantie des tests* $\hat{\pi}$, (5) *et* (6) *est égale à* $p_{c_\epsilon}(b)$.

DÉMONSTRATION. Le problème B consiste ici à tester au vu d'une observation $Y \in \Phi_{\gamma, E}$ l'hypothèse $\mathcal{R}_1 = \{|\gamma| \leqslant a\}$ contre $\mathcal{R}_2 = \{|\gamma| \geqslant b\}$. Nous avons vu dans l'exemple 9.1 que dans ce problème il existait un test minimax de niveau $1 - \epsilon$, de la forme

$$|Y| > c_\epsilon.$$

Nous nous sommes servis du théorème 9.1 pour construire ce test. Ceci exprime que les conditions du théorème 14.2 sont remplies. Donc, le test

$$|\gamma^*| > c_\epsilon$$

sera asymptotiquement minimax de niveau asymptotique $1-\epsilon$ pour le problème A.

Le test du rapport de vraisemblance (1) sera ici de la forme

$$R(X)=\frac{\sup\limits_{|u|\geqslant b} Z_1(un^{-1/2})}{\sup\limits_{|u|\leqslant a} Z_1(un^{-1/2})}>c. \tag{8}$$

En reprenant *ad litteram* les raisonnements des démonstrations des théorèmes 1 et 2, on trouve que $R(X)=T_n(X)+\epsilon_n(X), \epsilon_n(X)\underset{\mathbf{P}_{\theta_1}}{\to}0$, où

$$T_n(X)=r(\gamma^*)=\frac{\sup\limits_{|u|\geqslant b}\exp\left\{-\frac{1}{2}|\gamma^*-u|^2\right\}}{\sup\limits_{|u|\leqslant a}\exp\left\{-\frac{1}{2}|\gamma^*-u|^2\right\}}.$$

Comme dans ce qui précède, on déduit de là l'absolue continuité de la distribution de $r(Y)$ et l'équivalence asymptotique des tests $R(X)>c$ et $T_n(X)>c$. Le dernier test est équivalent au test

$$|\gamma^*|>c'$$

qui pour $c'=c_\epsilon$ sera un test de niveau $1-\epsilon$. En vertu du théorème 14.2 (cf. (14.10)) il admettra une puissance limite garantie égale à $p_{c_\epsilon}(b)$ (cf. théorème 9.2). ◄

REMARQUE 1. Si l'on revient au paramètre primitif, on constate que ce théorème est valable pour les hypothèses $H_i=\{\theta\in\Theta_i\}$, où (comparer avec l'exemple 9.2 pour $\sigma^2=I^{-1}$)

$$\Theta_1=\{\theta:(\theta-\theta_1)I(\theta_1)(\theta-\theta_1)^T\leqslant a^2n^{-1}\},$$
$$\Theta_2=\{\theta:(\theta-\theta_1)I(\theta_1)(\theta-\theta_1)^T\geqslant b^2n^{-1}\}.$$

Le test (6) devient

$$(\hat\theta^*-\theta_1)I(\theta_1)(\theta^*-\theta_1)^Tn>c_\epsilon^2$$

ou (cf. théorème 13.2)

$$L'(X,\theta_1)I^{-1}(\theta_1)(L'(X,\theta_1))^T>c_\epsilon^2. \tag{9}$$

Le test du rapport de vraisemblance ne change visiblement pas, puisque le maximum de $f_\theta(X)$ dans Θ_i est invariant par un changement de variables.

Signalons également que le test (9) est parfois de forme plus commode que (5) et (6), puisqu'il n'est pas lié au calcul de $\hat\theta^*$. On peut effectuer les mêmes changements pour les tests (2) et (3) dans les théorèmes 1 et 2. Nous laissons ceci au soin du lecteur.

REMARQUE 2. On peut construire exactement comme dans le théorème 3 un test asymptotiquement minimax pour des problèmes A susceptibles d'être ramenés au problème B envisagé dans l'exemple 9.5.

REMARQUE 3. Au § 13 nous avons construit un test asymptotiquement bayésien entre les hypothèses $\{\theta = \theta_1\}$ et $\{\theta \neq \theta_1\}$, ayant la forme du test du rapport de vraisemblance

$$\frac{f_{\theta^*}(X)}{f_{\theta_1}(X)} > c.$$

Ce test asymptotiquement bayésien est donc aussi asymptotiquement minimax entre les hypothèses $\{\theta = \theta_1\}$ et $\{(\theta - \theta_1)I(\theta_1)(\theta - \theta_1)^T \geqslant b^2 n^{-1}\}$ pour tout $b > 0$.

4. Test asymptotiquement minimax relatif à l'appartenance de la loi de l'échantillon à une sous-famille paramétrique. Nous allons étudier maintenant le test du rapport de vraisemblance dans un problème légèrement plus compliqué de choix entre les hypothèses $H_1 = \{\theta \in \Theta_1\}$ et $H_2 = \{\theta \in \Theta_2\}$ lorsque la dimension l de l'ensemble Θ_1 est telle que $0 < \dim \Theta_1 = l < k$, où $k > 1$. Plus exactement, soit donnée une fonction régulière $\theta = g(\alpha)$ d'un paramètre l-dimensionnel $(l < k)$ $\alpha \in A_1 \subset R^l$. Désignons par Θ_1 l'image de A_1 dans Θ par l'application g. Le problème consiste à choisir entre l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ que le paramètre θ appartient à la « courbe » Θ_1 (ou que $X \in \mathbf{P}_{g(\alpha)}$ pour un $\alpha \in A_1$) et son alternative $\{X \in \in \mathbf{P}_\theta ; \theta \notin \Theta_1\}$, de sorte que dans ce cas $\Theta_2 = \Theta \setminus \Theta_1$. En d'autres termes, ce problème consiste à vérifier que la loi de l'échantillon X appartient à une sous-famille paramétrique de distributions $\{\mathbf{P}_{g(\alpha)}, \alpha \in A_1\}$.

Font partie de cette classe les problèmes déjà envisagés de choix entre les hypothèses $\{X \in \Phi_{\alpha_0, \sigma^2}\}$ et $\{X \in \Phi_{\alpha, \sigma^2} ; \alpha \neq \alpha_0\}$ pour α_0 donné et σ^2 inconnue ou de choix entre les hypothèses $\{X \in \Phi_{\alpha, \sigma_0^2}\}$ et $\{X \in \Phi_{\alpha, \sigma^2} ; \sigma \neq \neq \sigma_0\}$ pour σ_0 donnée et α inconnue, etc.

On admettra que la courbe $\theta = g(\alpha)$ dans Θ est deux fois continûment différentiable et que la matrice $G = \| \partial g_i(\alpha)/\partial \alpha_j \|$ $(i = 1, \ldots, k ; j = 1, \ldots \ldots, l ;$ $g_i(\alpha)$ et α_i sont les coordonnées respectives de $g(\alpha)$ et $\alpha)$ est de rang l. Ceci exprime que nous pouvons effectuer un changement de paramètre biunivoque différentiable (une reparamétrisation du problème), de sorte que les l premières coordonnées (sans nuire à la généralité on peut les poser égales à $\alpha = (\alpha_1, \ldots, \alpha_l)$) définissent la position du point θ sur la courbe Θ_1 et les autres (que nous désignerons par $\beta = (\beta_1, \ldots, \beta_{k-l})$) la position du point θ dans le « plan » (dans le sous-espace), disons, orthogonal (mais pas nécessairement) à la « courbe » $g(\alpha)$ au point α. Le problème revient alors à choisir entre les hypothèses $\{\beta = 0\}$ et $\{\beta \neq 0\}$ en présence d'un sous-paramètre fantôme α inconnu.

Ceci étant, nous considérerons des hypothèses voisines en posant $\beta = \gamma'' n^{-1/2}$ et nous testerons l'hypothèse $\{\gamma'' = 0\}$ contre $\{\gamma'' \neq 0\}$ ou contre

$$\{\gamma'' M_2(\alpha)\gamma''^T \geqslant b^2\} \tag{10}$$

pour $b > 0$ et une matrice $M_2(\alpha)$ définie positive.

Dans les coordonnées primitives, le dernier problème consiste à tester l'hypothèse $H_1 = \{\theta \in \Theta_1\}$ contre des hypothèses concurrentes voisines où le paramètre θ est situé dans un $n^{-1/2}$-voisinage de la courbe Θ_1 et à l'extérieur d'un « tube » contenant Θ_1 et correspondant à l'ensemble (10). Une autre position du problème de test d'hypothèses voisines part du fait que le paramètre θ est « localisé » au voisinage d'un point $\theta_0 = g(\alpha^\circ)$, $\alpha^\circ \in A_1$. Le nouveau paramètre $\tau = (\beta, \alpha - \alpha^\circ)$ sera alors localisé au voisinage du point $\tau_0 = (0, 0)$. Posons $\alpha - \alpha^\circ = \gamma' n^{-1/2}$, $\beta = \gamma'' n^{-1/2}$ et éprouvons l'hypothèse $\{\gamma'' = 0\}$ contre $\{\gamma'' \neq 0\}$ ou contre $\{\gamma'' M_2(\alpha^\circ)\gamma''^T \geqslant b^2\}$ en présence d'un paramètre fantôme γ'.

Les résultats fournis par ces deux approches sont pratiquement les mêmes mais on optera pour la deuxième, car on dispose dans ce cas de toutes les données préliminaires nécessaires. L'hypothèse de la localisation du paramètre θ revêt un caractère conventionnel, et la forme des propositions établies plus bas sera indépendante de θ_0.

On admettra donc que le nouveau paramètre $\tau = (\alpha - \alpha^\circ, \beta)$ est de la forme

$$\tau = \gamma n^{-1/2}, \qquad \gamma = (\gamma', \gamma''),$$

et l'on éprouvera l'hypothèse $H_1 = \{\gamma'' = 0\}$ contre $H_2 = \{\gamma'' M_2 \gamma''^T \geqslant b^2\}$, où pour $M_2 = M_2(\alpha^\circ)$ on prendra la matrice d'information de Fisher pour la famille paramétrique $\{\mathbf{P}_{\theta(0,\beta)}\}$ au point $\beta = 0$, où $\theta(\tau) = \theta((\alpha - \alpha^\circ, \beta))$ est une fonction qui restitue θ au vu de $\tau = (\tau', \tau'')$.

THÉORÈME 4. *Supposons que $\theta_0 = g(\alpha^\circ)$ est un point intérieur de Θ au voisinage duquel sont remplies les conditions (RR). Supposons par ailleurs que la fonction $g(\alpha)$ est bicontinûment différentiable au point α° et que la matrice $G = \| \partial g_i(\alpha)/\partial \alpha_j \|_{\alpha=\alpha^\circ}$ est de rang l. Pour les ensembles Θ_1 et Θ_2 définis plus haut et pour un c convenable, le test du rapport de vraisemblance* (1) *est alors asymptotiquement équivalent aux tests*

$$R_1(X) \equiv \frac{f_{\hat\theta^*}(X)}{f_{g(\hat\alpha^*)}(X)} > e^{h_\epsilon/2}, \tag{11}$$

$$\begin{aligned} &(\hat\theta^* - g(\hat\alpha^*)) I(g(\hat\alpha^*))(\hat\theta^* - g(\hat\alpha^*))^T > h_\epsilon n^{-1}, \\ &(\hat\theta^* - g(\hat\alpha^*)) I(\hat\theta^*)(\hat\theta^* - g(\hat\alpha^*))^T > h_\epsilon n^{-1} \end{aligned} \tag{12}$$

et est un test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$ *de* $H_1 = \{\theta \in \Theta_1\} = \{\gamma'' = 0\}$ *contre* $H_2 = \{\gamma'' M_2 \gamma''^T \geqslant b^2\}$.

La distribution de la statistique $2 \ln R_1(X)$, *où* $X \in \mathbf{P}_{g(\alpha^\circ)}$ *(c'est-à-dire pour l'hypothèse* H_1*), converge, lorsque* $n \to \infty$, *vers une distribution du* χ^2 *à* $k - l$ *degrés de liberté (donc est indépendante de* f_θ *et* α°*). De ce fait, la quantité* k_ε *de* (11) *et* (12) *désigne le quantile d'ordre* $1 - \varepsilon$ *de la distribution* $\mathbf{H}_{k-l}$.

La puissance asymptotique garantie du test du rapport de vraisemblance est égale à $\mathbf{P}((\xi_1 + b)^2 + \xi_2^2 + \ldots + \xi_{k-l}^2 > k_\varepsilon)$, *où* $\xi_i \in \Phi_{0,1}$ *sont indépendantes.*

Nous voyons que les tests asymptotiquement minimax (11) et (12) ne dépendent en aucune façon de α°.

REMARQUE 4. Par rapport à θ l'hypothèse H_2 peut être mise sous la forme

$$H_2 = \{\inf_\gamma (\theta - g(\alpha^\circ + \gamma' n^{-1/2}) I(g(\alpha^\circ))(\theta - g(\alpha^\circ + \gamma' n^{-1/2}))^T \geqslant \geqslant b^2 n^{-1}\}.$$

On rappelle que l'on a postulé que les ensembles Γ_i sont bornés, de sorte que $(\theta - \theta_0) \leqslant N n^{-1/2}$, $|\gamma'| \leqslant N$ pour un $N > 0$.

REMARQUE 5. On verra dans la démonstration que le théorème reste entièrement en vigueur si l'on remplace $H_1 = \{\gamma'' = 0\}$ par $H_1 = \{\gamma'' M_2 \gamma''^T \leqslant a^2\}$, $a < b$, et l'ensemble Θ_1 par l'ensemble correspondant.

DÉMONSTRATION du théorème 4. Pour « principal » test on prendra le test (11) qui est équivalent à (1) et de forme plus commode. Nous montrerons qu'il est asymptotiquement équivalent à un test asymptotiquement minimax et ensuite qu'il est asymptotiquement équivalent au test (12).

Traitons les distributions $\mathbf{P}_\theta$ et $\mathbf{P}_{g(\alpha)}$ comme des distributions dépendant des paramètres $\tau = (\tau', \tau'')$ et $\alpha = \tau' + \alpha^\circ$ respectivement. Posons $\tau = \gamma n^{-1/2}$, $\gamma = (\gamma', \gamma'')$, de sorte que $\tau' = \gamma' n^{-1/2}$, $\tau'' = \gamma'' n^{-1/2}$, et testons l'hypothèse $H_1 = \{\gamma'' = 0\}$ contre $H_2 = \{\gamma'' M_2 \gamma''^T \geqslant b^2\}$, où M_2 est la matrice d'information de Fisher pour la famille $\{\mathbf{P}_{g(\alpha)}\}$ au point α°. Effectuons maintenant une autre transformation sur le paramètre θ comme nous l'avons fait dans l'exemple 9.4 pour transformer les matrices d'information en matrices unités. Plus exactement, posons $\rho = \tau\Lambda$ et respectivement $\delta = \gamma\Lambda$ ($\rho = \delta n^{-1/2}$), où Λ est une matrice triangulaire décrite en détail dans l'exemple 9.4 et douée des propriétés suivantes :

$$J^{-1} = \Lambda^T M^{-1} \Lambda = E, \quad J_2^{-1} = \Lambda_2^T M_2^{-1} \Lambda_2 = E,$$

où J, M, J_2 et M_2 sont les matrices d'information au point θ_0 respectivement pour ρ, τ, ρ'' et τ'' (les accents ont la même signification que ceux de τ', τ'', γ' et γ''), Λ_2 est une $(k - l)$-matrice formée par les $k - l$ dernières lignes et colonnes de la matrice Λ, de sorte que $\rho'' = \tau''\Lambda_2$, $\delta'' = \gamma''\Lambda_2$.

Par rapport aux nouveaux paramètres, les hypothèses H_1 et H_2 s'écrivent

$$H_1 = \{\delta'' = 0\}, \quad H_2 = \{|\delta''| \geqslant b\}.$$

Les propriétés des transformations effectuées nous montrent clairement que $\theta = \theta(\rho)$ est une application bijective et que toutes les familles paramétriques envisagées (y compris celles

dépendant des paramètres ρ et ρ'') vérifient les conditions (RR). Posons $\rho_0 = \theta^{-1}(\theta_0)$ (ceci est la solution de l'équation $\theta(\rho) = \theta_0$)

$$Z_0(t) = f_{\theta(\rho_0+t)}(X)/f_{\theta_0}(X), \quad Y_0(u) = \ln Z_0(un^{-1/2}).$$

Utilisons le théorème 2.29.3. Nous obtenons pour $|u| \leqslant \delta_n\sqrt{n}$, $X \in \mathbf{P}_{\theta(\rho)}$

$$\rho = \rho_0 + \delta n^{-1/2},$$

$$Y_0(u) = (\xi_n + \delta, u) - \frac{1}{2}(u,u) + (|u|^2 + |\delta|^2)\epsilon_n(X,u,\delta), \tag{13}$$

où $|\epsilon_n(X,u,\delta)| \leqslant \epsilon_n(X) \underset{P_{\theta(\rho)}}{\to} 0$ uniformément en δ pour $|\delta| \leqslant \delta_n\sqrt{n}$, δ_n étant une suite arbitraire convergeant vers 0. Dans ces égalités on s'est servi du fait que la matrice d'information pour le paramètre ρ est une matrice unité. Le vecteur ξ_n est le vecteur des dérivées de la fonction $n^{-1/2}L(X,\theta(\rho))$ par rapport à ρ_j au point $\rho = \rho_0 + \delta n^{-1/2}$, de sorte que $\xi_n \in \Phi_{0,E}$ uniformément en ρ (en δ) pour $|\delta| \leqslant \delta_n\sqrt{n}$. (Vu que nous avons admis que $(\theta - \theta_0)\sqrt{n}$ était borné, il suffit dans la suite d'établir la convergence uniforme dans $|\delta| \leqslant N$ pour un N fixe quelconque. Mais rien ne nous empêche d'établir l'uniformité exigée dans le domaine plus vaste $|\delta| \leqslant \delta_n\sqrt{n} \to \infty$.)

Posons maintenant $u = (u', u'')$, $u'' = 0$, dans (13). On a alors, compte tenu de la convention précédente sur les notations avec les accents,

$$Y_0((u',0)) = (\xi_n' + \delta', u') - \frac{1}{2}(u',u') + (|u'|^2 + |\delta|^2)\epsilon_n(X,u',\delta). \tag{14}$$

Des relations (13) et (14) on déduit que $Y_0(u)$ et $Y_0((u',0))$ atteignent leurs maximums respectivement pour

$$\begin{aligned} u &= (\xi_n + \delta)(E + \epsilon_n(X,\delta)), \\ u' &= (\xi_n' + \delta')(E + \epsilon_n^{(1)}(X,\delta)), \end{aligned} \tag{15}$$

où $\epsilon_n(X,\delta) \underset{P_{\theta(\rho)}}{\to} 0$, $\epsilon_n^{(1)}(X,\delta) \underset{P_{\theta(\rho)}}{\to} 0$ uniformément en δ, $|\delta| < \delta_n\sqrt{n}/2$. Il suffit simplement de remarquer que la probabilité des grandes valeurs $|\xi_n + \delta|$ est uniformément petite, puisque $\xi_n + \delta \in \Phi_{\delta,E}$ uniformément en δ, $|\delta| < \delta_n\sqrt{n}$, et $\mathbf{P}_\theta(|\xi_n + \delta| > \delta_n\sqrt{n}) \to 0$ uniformément en δ, $|\delta| < \delta_n\sqrt{n}/2$.

Considérons maintenant le test du rapport de vraisemblance. Pour $\theta = \theta(\rho)$, $X \in \mathbf{P}_\theta$, $\rho = \rho_0 + \delta n^{-1/2}$, on a

$$R_1(X) = \frac{\sup_\theta f_\theta(X)}{\sup_\alpha f_{g(\alpha)}(X)} = \frac{\sup_u e^{Y_0(u)}}{\sup_{u'} e^{Y_0((u',0))}} =$$

$$= \frac{\exp\left\{\frac{1}{2}|\xi_n + \delta|^2 + \bar{\epsilon}_n(X,\delta)\right\}}{\exp\left\{\frac{1}{2}|\xi_n' + \delta'|^2 + \bar{\epsilon}_n^{(1)}(X,\delta)\right\}} = \exp\left\{\frac{1}{2}|\xi_n'' + \delta''|^2 + \epsilon_n''(X,\delta)\right\}, \tag{16}$$

où les fonctions ϵ_n affectées d'indices différents convergent vers 0 en $\mathbf{P}_\theta$-probabilité uniformément pour $|\delta| < \delta_n\sqrt{n}$;

$$2 \ln R_1(X) \Rightarrow |Y'' + \delta''|^2, \quad Y \in \Phi_{0,E}, \tag{17}$$

uniformément en δ.

Etant donné que pour $\theta = g(\alpha)$ on a nécessairement $\delta'' = 0$, on en déduit l'assertion du théorème relativement à la statistique $2 \ln R_1(X)$.

Si on se rappelle maintenant (cf. théorème 2.29.3) que $\xi_n = u^*(E + \epsilon_n(X, \delta))$, où $u^* = (\hat{\rho}^* - \rho)\sqrt{n}$ et $\hat{\rho}^*$ est un estimateur du maximum de vraisemblance de ρ, on déduit de là et de l'égalité $\rho_0'' = 0$, en posant $\delta^* = (\hat{\rho}^* - \rho_0)\sqrt{n}$, que

$$\xi_n + \delta = \sqrt{n}(\hat{\rho}^* - \rho + \rho - \rho_0) + u^*\epsilon_n(X, \delta) = \sqrt{n}(\hat{\rho}^* - \rho_0) + u^*\epsilon_n(X, \delta) = \delta^* + u^*\epsilon_n(X, \delta) \in \Phi_{\delta, E},$$

$$\xi_n'' + \delta'' = (\delta^*)'' + (u^*\epsilon_n(X, \delta))''.$$

Donc, le membre de droite de (16) peut être mis aussi sous la forme $\exp\left\{\frac{1}{2}|(\delta^*)''|^2 + \epsilon_n'''(X, \delta)\right\}$, $\epsilon_n'''(X, \delta) \underset{P_\theta}{\to} 0$. Ce qui exprime que le test

$$|(\delta^*)''|^2 > h_\epsilon \tag{18}$$

est asymptotiquement équivalent à un test du rapport de vraisemblance, c'est-à-dire que

$$\lim_{n\to\infty} \sup_\alpha \mathbf{P}_{g(\alpha)}(R_1(X) > e^{h_\epsilon/2}) = \lim_{n\to\infty} \sup_\alpha \mathbf{P}_{g(\alpha)}(|(\delta^*)''| > h_\epsilon) = \epsilon,$$

$$\lim_{n\to\infty} \sup_{\theta\in\Theta_2} \mathbf{P}_\theta(R_1(X) > e^{h_\epsilon/2}) = \lim_{n\to\infty} \sup_{\theta\in\Theta_2} \mathbf{P}_\theta(|(\delta^*)''|^2 > h_\epsilon) =$$

$$= \sup_{|\delta''|>b^2} \mathbf{P}(|Y'' + \delta''|^2 > h_\epsilon) = \mathbf{P}((y_1 + b)^2 + y_2^2 + \ldots + y_{k-l}^2 > h_\epsilon),$$

où $y_i \in \Phi_{0,1}$ sont indépendantes.

Montrons maintenant que le test (18) est un test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$. Utilisons le théorème 14.2. Dans notre cas $\delta^* = (\hat{\rho}^* - \rho_0)\sqrt{n} \in \Phi_{\delta, E}$. Nous avons étudié le problème B pour $Y \in \Phi_{\delta, E}$ dans les exemples 9.3 et 9.4. Nous avons établi que le test

$$|Y''|^2 > h_\epsilon$$

est minimax de niveau $1 - \epsilon$. Donc, le test (18) est asymptotiquement minimax en vertu du théorème 14.2.

Pour achever la démonstration il reste à établir l'équivalence asymptotique de (11) et de (12). Cette équivalence découle sans peine des résultats du § 2.29 et du lemme 14.1. ◄

EXEMPLE 1. Soit $X \in \Phi_{\lambda, \sigma^2}$, où λ et σ^2 sont des paramètres scalaires. (Nous utilisons ici λ au lieu du traditionnel α pour éviter toute confusion avec l'argument de la fonction $g(\alpha)$.) On demande de tester l'hypothèse $\{\lambda = \lambda_0\}$ contre $\{\lambda \neq \lambda_0\}$ ou contre $\{|\lambda - \lambda_0| \geq bn^{-1/2}\}$, $b > 0$, le paramètre σ étant inconnu. Si les composantes λ et σ^2 du vecteur $\theta = (\lambda, \sigma^2)$ sont toutes deux inconnues, un estimateur du maximum de vraisemblance pour

θ est

$$\hat{\theta}^* = (\lambda, \sigma^2)^* = (\bar{x}, S^2), \quad S^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

Si $\lambda = \lambda_0$, un estimateur du maximum de vraisemblance pour σ^2 est : $(\sigma^2)^* = S_1^2 = \frac{1}{n} \sum (x_i - \lambda_0)^2$, de sorte que $g(\hat{\alpha}^*) = (\lambda_0, S_1^2)$. Comme

$$f_\theta(X) = (\sqrt{2\pi}\sigma)^{-n} \exp\{-(2\sigma^2)^{-1} \sum (x_i - \lambda)^2\},$$

le test du rapport de vraisemblance (11) sera de la forme

$$S_1^2/S^2 > c.$$

Puisque $S_1^2 = S^2 + (\bar{x} - \lambda_0)^2$, ce test est équivalent à

$$|\bar{x} - \lambda_0|/S > c_1. \tag{19}$$

On reconnaît ici le test classique de Student que nous avons étudié antérieurement (les propriétés optimales de ce test sont accessibles au § 7).

Il est immédiat de vérifier que le test (12) sera de la même forme. En effet, au § 2.16 nous avons vu que la matrice $I(\theta)$ pour la famille Φ_{λ, σ^2} est de la forme

$$I(\theta) = \begin{pmatrix} \sigma^{-2} & 0 \\ 0 & (2\sigma^4)^{-1} \end{pmatrix}.$$

Ici $\hat{\theta}^* - g(\hat{\alpha}^*) = (\bar{x} - \lambda_0, S^2 - S_1^2) = (\bar{x} - \lambda_0, n(\bar{x} - \lambda_0)^2)$,

$$I^{1/2}(\hat{\theta}^*) = \begin{pmatrix} S^{-1} & 0 \\ 0 & (\sqrt{2}S^2)^{-1} \end{pmatrix}.$$

Puisque le premier membre de (12) est le carré de la norme $|(g(\hat{\alpha}^*) - \hat{\theta}^*)I^{1/2}(\hat{\theta}^*)|^2$, le test (12) sera de la forme

$$\frac{(\bar{x} - \lambda_0)^2}{S^2} + \frac{(\bar{x} - \lambda_0)^4}{2S^4} > c_2,$$

qui est visiblement équivalente à (19).

Si on utilise $I(g(\hat{\alpha}^*))$ au lieu de $I(\hat{\theta}^*)$, on obtient le test asymptotiquement équivalent

$$|\bar{x} - \lambda_0|/S_1 > c_1.$$

EXEMPLE 2. Soit $X \in \Phi_{\lambda, \sigma^2}$. On demande de tester l'hypothèse $\{\sigma = \sigma_0\}$ contre $\{|\sigma^2 - \sigma_0^2| \geqslant bn^{-1/2}\}$ lorsque λ est inconnu. Il est évident que l'estimateur du maximum de vraisemblance $\hat{\theta}^*$ de $\theta = (\lambda, \sigma^2)$ sera le même que dans l'exemple précédent. Si $\sigma = \sigma_0$, alors $\hat{\lambda}^* = \bar{x}$, et $g(\hat{\alpha}^*) = (\bar{x}, \sigma_0^2)$, $\hat{\theta}^* - g(\hat{\alpha}^*) = (0, \sigma_0^2 - S^2)$.

Les tests (11) (ou ce qui revient au même le test du rapport de vraisemblance) seront de la forme

$$(S^2 - \sigma_0^2)^2/\sigma_0^4 > 2h_\epsilon n^{-1},$$

qui est visiblement équivalente à

$$|S^2/\sigma_0^2 - 1| > \sqrt{2h_\epsilon n^{-1}},$$

où $\Phi_{0,1}(]h_\epsilon^{1/2}, \infty[) = \epsilon/2$. Nous avons étudié ce test aussi au § 7.

D'autres exemples d'application du théorème 4 sont accessibles au § 17.

§ 16. Test du χ^2. Test d'hypothèses d'après des données groupées

1. Test du χ^2. Propriétés d'optimalité asymptotique. Le test du χ^2 est « initialement » un test de choix entre l'hypothèse simple $H_1 = \{\theta = p\}$ et sa complémentaire $H_2 = \{\theta \neq p\}, p = (p_1, ..., p_r)$, au vu d'un échantillon X issu de la loi polynomiale $\mathbf{B}_\theta, \theta = (\theta_1, ..., \theta_r), \sum_{i=1}^{r} \theta_i = 1$. La distribution polynomiale $\mathbf{B}_\theta$ est décrite par les probabilités $\theta_i = \mathbf{P}(A_i), i = 1, ..., r$, d'apparition en une épreuve de l'un des r événements disjoints $A_1, ..., A_r$. On peut se représenter un élément x_j de $X \in \mathbf{B}_\theta$ comme un vecteur $e_k, k = 1, ..., r$, dont la composante d'indice k est égale à 1 et les autres à 0 ; de plus $x_j = e_k$ si l'événement A_k s'est produit. Désignons par ν_k le nombre d'apparitions de A_k en n épreuves indépendantes. Alors $\nu = (\nu_1, ..., \nu_r) = \sum_{i=1}^{n} x_i$ est une statistique exhaustive pour θ, puisque la fonction de vraisemblance $f_\theta(X)$ est de la forme

$$f_\theta(X) = \prod_{i=1}^{r} \theta_i^{\nu_i}. \tag{1}$$

La statistique χ^2 est par définition

$$\chi^2(X) = \sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i},$$

et la région critique du test du χ^2 (la région d'acceptation de H_2) est de la forme

$$\chi^2(X) \geqslant c,$$

où c est choisi en fonction du niveau du test.

Appesantissons-nous sur le problème de test de $H_1 = \{\theta = p\}$ contre $H_2 = \{\theta \neq p\}$.

Il est clair que les distributions $\mathbf{B}_\theta$ forment une famille paramétrique dépendant d'un paramètre $(\theta_1, ..., \theta_{r-1})$ de dimension $k = r - 1$; le paramètre θ_r prend la valeur $\theta_r = 1 - \sum_{i=1}^{r-1} \theta_i$. On désignera les vecteurs $(\theta_1, ...$ $..., \theta_{r-1})$ et $(\theta_1, ..., \theta_r)$ par la même lettre θ sans crainte de confusion. La région Θ est le simplexe $\theta_i \geqslant 0, i = 1, .., r-1, \sum_{i=1}^{r-1} \theta_i \leqslant 1$. Le logarithme de la fonction de vraisemblance $L(X, \theta)$ est égal à

$$L(X,\theta) = \sum_{k=1}^{r} \nu_k \ln \theta_k = \sum_{i=1}^{n} l(\mathrm{x}_i, \theta). \tag{2}$$

La famille $\{\mathbf{B}_\theta\}$ vérifie les conditions (A_0), (A_μ), (A_c) ainsi que les conditions de régularité (RR) en tout point intérieur de Θ, c'est-à-dire en tout point θ pour lequel tous les $\theta_i > 0$. En effet, dans le cas considéré

$$l(\mathrm{x}_1, \theta) = \ln \theta_j \quad \text{si} \quad \mathrm{x}_1 = e_j ;$$

$$\frac{\partial l(\mathrm{x}_1,\theta)}{\partial \theta_j} = \begin{cases} \theta_j^{-1} & \text{si} \quad \mathrm{x}_1 = e_j, \\ -\theta_r^{-1} & \text{si} \quad \mathrm{x}_1 = e_r, \\ 0 & \text{si} \quad \mathrm{x}_1 \neq e_j, \quad \mathrm{x}_1 \neq e_r, \end{cases} \tag{3}$$

$$\frac{\partial^2 l(\mathrm{x}_1,\theta)}{\partial \theta_i \partial \theta_j} = \begin{cases} -\dfrac{\delta_{ij}}{\theta_i \theta_j} & \text{si} \quad \mathrm{x}_1 = e_j, \\ -\theta_r^{-2} & \text{si} \quad \mathrm{x}_1 = e_r, \\ 0 & \text{si} \quad \mathrm{x}_1 \neq e_j, \mathrm{x}_1 \neq e_r, \end{cases} \tag{4}$$

où δ_{ij} est le symbole de Kronecker. On voit sur ces formules que

$$\frac{\partial^2 l(\mathrm{x}_1,\theta)}{\partial \theta_i \partial \theta_j} = - \frac{\partial l(\mathrm{x}_1,\theta)}{\partial \theta_i} \cdot \frac{\partial l(\mathrm{x}_1,\theta)}{\partial \theta_j}, \quad i,j \leqslant r-1.$$

La partie des conditions (RR) concernant l'existence des espérances mathématiques est manifestement réalisée, puisque l'ensemble $\mathscr{X}$ est fini.

De (3) et (4), on déduit que

$$I(\theta) = \| I_{ij}(\theta) \| = - \left\| \mathbf{E}_\theta \frac{\partial^2 l(\mathrm{x}_1,\theta)}{\partial \theta_i \partial \theta_j} \right\| = \left\| \frac{\delta_{ij}}{\theta_i} + \frac{1}{\theta_r} \right\|, \tag{5}$$

$$i,j = 1, ..., r-1.$$

Dans cette matrice, si l'on soustrait la première ligne de toutes les autres et que l'on développe le déterminant suivant cette première ligne, on obtient

$$|I(\theta)| = \left(1 + \sum_{j=1}^{r-1} \frac{\theta_j}{\theta_r}\right) \prod_{j=1}^{r-1} \theta_j^{-1} = \left(\prod_{j=1}^{r} \theta_j\right)^{-1}.$$

Donc, $0 < |I(\theta)| < \infty$ si $\prod_{k=1}^{r} \theta_k > 0$, c'est-à-dire si θ est un point intérieur du simplexe Θ.

Nous voyons donc que nous pouvons à juste titre appliquer les résultats des §§ 13 et 14 sur les tests asymptotiquement optimaux. De ces résultats il découle qu'il existe un test asymptotiquement bayésien de $H_1 = \{\theta = p\}$ contre $H_2 = \{\theta \neq p\}$ qui est confondu avec le test du rapport de vraisemblance

$$\frac{f_{\hat{\theta}^*}(X)}{f_p(X)} > c. \tag{6}$$

Ce même test sera asymptotiquement minimax de H_1 contre $\{(\theta - p)I(\theta)(\theta - p)^T > b^2 n^{-1}\}$ (cf. théorème 15.3).

Pour déterminer la région critique de (6) sous une forme plus commode, calculons la valeur $f_{\hat{\theta}^*}(X)$. Une dérivation de (2) par rapport à $\theta_1, \ldots, \theta_{r-1}$ nous donne

$$\frac{\partial L(X, \theta)}{\partial \theta_i} = \frac{\nu_i}{\theta_i} - \frac{\nu_r}{\theta_r}, \quad i = 1, \ldots, r-1.$$

En égalant ces dérivées à zéro, on trouve que l'estimateur du maximum de vraisemblance est

$$\hat{\theta}^* = n^{-1}\nu,$$

de sorte que $\hat{\theta}_i^* = n^{-1}\nu_i$.

En passant aux logarithmes, on peut donc représenter le test (6) sous la forme

$$\psi^2(X) = \sum_{i=1}^{r} \nu_i \ln \frac{\nu_i}{np_i} > c_1. \tag{7}$$

Le théorème 13.1 (cf. aussi le lemme 13.1) nous dit que la statistique $2\psi^2(X)$ pour l'hypothèse H_1 admet une distribution limite du χ^2 à $r-1$ degrés de liberté. Pour cette raison, on obtient un test de niveau asymptotique $1-\epsilon$ si l'on pose $c_1 = h_\epsilon/2$, où h_ϵ est le quantile d'ordre $1-\epsilon$ de la distribution $\mathbf{H}_{r-1}$.

Comment se présente dans nos conditions le test π' asymptotiquement équivalent à (6), obtenu dans le théorème 13.2 sous la forme

$$n(\hat{\theta}^* - p)I(p)(\hat{\theta}^* - p)^T > h_\epsilon \,? \tag{8}$$

Pour $t = (t_1, \ldots, t_{r-1}), s = \sum_{i=1}^{r-1} t_i$, on a

$$tI(p) = \left(\frac{t_1}{p_1} + \frac{s}{p_r}, \ldots, \frac{t_{r-1}}{p_{r-1}} + \frac{s}{p_r}\right),$$

$$tI(p)t^T = \sum_{i=1}^{r-1} \frac{t_i^2}{p_i} + \frac{s^2}{p_r} = \sum_{i=1}^{r} \frac{t_i^2}{p_i}, \tag{9}$$

où

$$t_r = -s, \quad \sum_{i=1}^{r} t_i = 0. \tag{10}$$

En posant $t = \hat{\theta}^* - p$ et en remarquant que la condition (10) est remplie, on obtient pour (8) le test

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{np_i} > h_\epsilon. \tag{11}$$

Ce qui n'est autre qu'un test du χ^2. Des propositions précédentes il résulte que $\chi^2(X) \in \mathbf{H}_{r-1}$.

Le test π'' du théorème 13.2, qui est asymptotiquement équivalent à (7) et à (11), sera de la forme

$$\sum_{i=1}^{r} \frac{(\nu_i - np_i)^2}{\nu_i} > h_\epsilon. \tag{12}$$

En tenant compte aussi du théorème 15.3 et de la remarque 15.1, on peut résumer ce qui vient d'être dit par la proposition suivante.

THÉORÈME 1. *Le test* (7) *pour* $c_1 = h_\epsilon/2$ *ainsi que le test du* χ^2 (11) *et le test* (12) *possèdent un niveau asymptotique égal à* $1 - \epsilon$ *et sont des tests asymptotiquement bayésiens de* $\{\theta = p\}$ *contre* $\{\theta \neq p\}$ *au vu de* $X \in \mathbf{B}_\theta$. *Ce sont également des tests asymptotiquement minimax de* $\{\theta = p\}$ *contre l'alternative voisine* $\left\{\sum_{i=1}^{r} (\theta_i - p_i)^2/p_i > b^2/n\right\}$ *pour tout* $b > 0$.

On aurait pu établir directement l'équivalence asymptotique des tests (7), (11) et (12) en développant $\ln \frac{\nu_i}{np_i} = \ln\left(1 + \frac{\nu_i - np_i}{np_i}\right)$ en série dans (7).

Ces tests sont asymptotiquement non paramétriques, puisque la distribution limite des statistiques utilisées est « absolue », c'est-à-dire n'est pas liée à la nature de la distribution initiale.

2. Applications du test du χ^2. Test d'hypothèses d'après des données groupées. Le test du χ^2 est très répandu et sa portée dépasse le cadre du problème envisagé dans le numéro précédent.

Considérons le problème général de choix entre l'hypothèse $H_1 = \{X \in \in \mathbf{P}_1\}$ et l'hypothèse $H_2 = \{X \in \mathbf{P}, \mathbf{P} \neq \mathbf{P}_1\}$, étudié dans le § 12. Vu qu'une théorie des tests optimaux, tant soit peu développée, n'existe que pour le cas paramétrique, une idée assez naturelle est de tenter de « paramétrer » *) ce problème d'une manière ou d'une autre.

Le moyen le plus simple et le plus naturel dans le cas général est le *groupement des données* qui consiste en ce qui suit. On divise le domaine des valeurs possibles des variables observées (c'est-à-dire l'espace $\mathscr{X}$) en r domaines disjoints $\Delta_1, \ldots, \Delta_r$, et au lieu de l'observation x_j on n'envisage que l'intervalle Δ_k la contenant. En d'autres termes, nous rendons les observations plus grossières, et les x_j contenus dans Δ_k, nous pouvons les remplacer par une valeur $z_k \in \Delta_k$. Il est clair qu'en choisissant une partition assez fine on peut approcher x_j d'aussi près que l'on veut par z_k.

Ainsi, le groupement conduit à remplacer l'observation x_j par un vecteur e_k si l'événement $A_k = \{\mathrm{x}_j \in \Delta_k\}$ s'est produit (les vecteurs e_k ont été définis au début du numéro précédent). Mais le nouvel échantillon obtenu par cette procédure n'est autre, de toute évidence, qu'un échantillon distribué suivant la loi $\mathbf{B}_\theta$, $\theta_k = \mathbf{P}(\mathrm{x}_j \in \Delta_k)$. Nous savons déjà que dans ce cas le vecteur $\nu = (\nu_1, \ldots, \nu_r)$ des fréquences d'atteinte des intervalles $\Delta_1, \ldots, \Delta_r$ sera une statistique exhaustive.

Cette réduction de l'échantillon X au vecteur ν s'appelle groupement des données.

Il est clair que le groupement se solde par un certain « appauvrissement » de l'échantillon X et par une perte partielle d'information.

Mais cette paramétrisation peut être envisagée sous un angle légèrement différent. Supposons pour plus de suggestion que $\mathscr{X} = R$ et que toutes les

*) On a en vue un paramètre θ de dimension finie. Tout problème peut être considéré comme étant paramétrique si l'on admet que θ est de dimension infinie, car on peut alors l'identifier à $\mathbf{P}$, $X \in \mathbf{P}$.

distributions envisagées sont concentrées dans un intervalle fini et admettent une densité, c'est-à-dire vérifient la condition (A_μ), où μ est la mesure de Lebesgue. Pour une subdivision donnée $\{\Delta_1, \ldots, \Delta_r\}$, considérons la densité $f(x)$ et la densité constante par morceaux

$$f_\theta(x) = \frac{\mathbf{P}(\Delta_i)}{\Delta_i} = \frac{1}{\Delta_i} \int_{\Delta_i} f(x)dx = \frac{\theta_i}{\Delta_i} \quad \text{pour} \quad x \in \Delta_i. \tag{13}$$

On désigne aussi par Δ_i la longueur de l'intervalle Δ_i. On reconnaît ici une famille paramétrique de distributions $\mathbf{P}_\theta$, $\mathbf{P}_\theta(B) = \int_B f_\theta(x)dx$.

On obtient un échantillon $Y \in \mathbf{P}_\theta$ si pour chaque k on rassemble toutes les observations de $X \in \mathbf{P}$ tombant dans Δ_k et qu'on les répartisse de façon aléatoire et uniforme sur Δ_k. Au fond nous avons réalisé la même chose que précédemment, puisque le fait de savoir en quel point de l'intervalle Δ_k se trouve l'observation y_i ne fournit aucune information sur le paramètre θ : la fonction de vraisemblance $f_\theta(Y)$ ne change pas dans les limites de ses intervalles lorsqu'on « déplace » les observations. Il suffit donc de connaître le nombre $\nu_1, \ldots, \nu_r$ des observations contenues respectivement dans $\Delta_1, \ldots, \Delta_r$.

Il est clair que si $f(x)$ est une fonction régulière, $f_\theta(x)$ l'approchera assez bien pour une subdivision $\{\Delta_1, \ldots, \Delta_r\}$ assez fine.

Les relations (13) définissent un autre procédé de paramétrisation qui est équivalent au premier. L'équivalence résulte de la coïncidence, à un facteur multiplicatif près indépendant du paramètre, des fonctions de vraisemblance. Pour la distribution (13), la fonction de vraisemblance est égale à

$$f_\theta(Y) = \prod_{i=1}^{r} \theta_i^{\nu_i} \prod_{i=1}^{r} \Delta_i^{-\nu_i},$$

où le premier facteur est la fonction de vraisemblance pour un échantillon distribué selon la loi $\mathbf{B}_\theta$ (cf. (1)).

A noter que le groupement des observations se présente assez fréquemment en soi non pas à des fins de paramétrisation, mais simplement comme un procédé plus commode et plus économique de représentation de l'information contenue dans un échantillon. Si par exemple $n = 10^4$ et que les valeurs observées sur l'intervalle [0, 1] soient mesurées au dixième près, il est clair alors qu'il est pratiquement superflue de connaître toutes les 10^4 observations et qu'il suffit d'indiquer les 10 fréquences $\nu_1, \ldots, \nu_{10}$ d'accès aux intervalles $\Delta_i =](i-1)/10, i/10[$, $i = 1, \ldots, 10$, c'est-à-dire de connaître seulement l'*histogramme de l'échantillon*.

Revenons au problème de choix entre $H_1 = \{X \in \mathbf{P}_1\}$ et $H_2 = \{X \in \mathbf{P} \neq \neq \mathbf{P}_1\}$. On admettra que le groupement des observations est tel que le désaccord entre $\mathbf{P}$ et $\mathbf{P}_1$, qui est significatif pour nous, se répercutera obligatoirement sur la distribution des données groupées. Notre problème peut alors être considéré comme un problème de choix entre l'hypothèse $\{\theta = = p\}$, où $p_i = \mathbf{P}_1(\Delta_i)$ et l'hypothèse contraire $\{\theta \neq p\}$ pour les familles paramétriques $\mathbf{B}_\theta$ ou (13). On sait déjà que *le test du* χ^2 (*de même que les tests* (7) *et* (12)) *sera asymptotiquement optimal dans ce problème au sens formulé dans le théorème* 1.

Par ailleurs, le test du χ^2 est *asymptotiquement non paramétrique*, puisque la distribution limite de la statistique $\chi^2(X)$ pour H_1 ne dépend pas de la distribution initiale de l'échantillon X.

Ceci étant, signalons que le test de l'hypothèse $\{\theta = p\}$ pour les familles (13) ou $\mathbf{B}_\theta$ n'est pourtant pas équivalent au test de l'hypothèse $\{X \in \mathbf{P}_1\}$, bien qu'il puisse en être proche si la subdivision $\{\Delta_1, \ldots, \Delta_r\}$ est assez fine. En effet, on teste l'hypothèse $X \in \mathbf{P}, \mathbf{P}(\Delta_i) = p_i = \mathbf{P}_1(\Delta_i)$. Ceci rend le test du χ^2 *non convergent* par rapport aux hypothèses alternatives $\mathbf{P} \neq \mathbf{P}_1$ telles que $\theta_i = \mathbf{P}(\Delta_i) = \mathbf{P}_1(\Delta_i) = p_i$. Nous pouvons donc noter une fois de plus que le test du χ^2 est un test qui est doué de nombreuses propriétés d'optimalité asymptotique mais qui n'agit que contre les alternatives modifiant le vecteur θ, c'est-à-dire contre les alternatives pour lesquelles $\{\mathbf{P}(\Delta_i)\} \neq \neq \{\mathbf{P}_1(\Delta_i)\} = \{p_i\}$.

Faisons quelques remarques sur les applications des tests du χ^2, (7) et (12). On parlera essentiellement du test du χ^2, puisque d'une part les tests précités sont voisins l'un de l'autre, et d'autre part le test du χ^2 (en partie en raison de sa suggestivité) est de loin le plus répandu.

Le niveau du test du $\chi^2(X) > h_\epsilon$ n'est égal à $1 - \epsilon$ qu'à la « limite ». L'expérience montre que pour $\epsilon \geqslant 0{,}01$ le vrai niveau n'est passablement approché par $1 - \epsilon$ que pour $np_i \geqslant 8$, $i = 1, \ldots, r$.

Lorsque le nombre r de groupes est élevé, disons que $n > r > 30$, on peut se servir de l'approximation normale aussi bien pour la distribution de $\frac{1}{\sqrt{2r}}(\chi_r^2 - r)$, $\chi_r^2 \in \mathbf{H}_r$ (cf. § 2.2) que pour la distribution pour H_1 de la statistique $\chi^2(X)$ normée par les moments

$$\mathsf{E}\chi^2(X) = r - 1,$$

$$\mathsf{V}\chi^2(X) = 2(r-1) + \frac{1}{n}\left(\sum_{i=1}^{r} p_i^{-1} - r^2 - 2r + 2\right).$$

On se sert souvent aussi de l'approximation normale $\Phi_{0,1}$ pour la distribution de la variable aléatoire (cf. § 2.2) $\sqrt{2\chi_r^2} - \sqrt{2r-1}$, $\chi_r^2 \in \mathbf{H}_r$.

Signalons également que lorsque le nombre de groupes croît, la densité $f(x)$ se laisse mieux approcher par une fonction en escalier construite à l'aide des valeurs $\mathbf{P}_1(\Delta_i) = \int_{\Delta_i} f(x)dx$. Ceci exprime que le nombre d'hypothèses contraires à H_1 croît, et le test du χ^2 a tendance à devenir un test relatif à la *densité*. Donc, la puissance d'un test du χ^2 de niveau fixé diminuera lorsque le nombre de groupes augmentera (comparer avec les remarques du paragraphe précédent sur le test de Moran. Pour plus de détails voir [12], [81]).

Au chapitre des défauts du test du χ^2 il faut noter que dans bien des cas c'est au statisticien de choisir la subdivision $\{\Delta_1, ..., \Delta_r\}$. Une certaine prudence est à conseiller, car l'« appauvrissement » de l'échantillon est réalisé de façon subjective. Par ailleurs, cette subdivision est choisie parfois en fonction de l'échantillon X, ce qui n'est pas toujours toléré, puisque les Δ_i deviennent alors aléatoires (pour plus de détails voir [43]).

EXEMPLE 1 *). Dans une ville N on a relevé l'heure indiquée par 500 montres exposées dans les vitrines de diverses horlogeries. Les résultats des observations ont été répartis en 12 groupes (en fonction de la position de l'aiguille des heures) dans le tableau suivant :

Secteurs horaires	0—1	1—2	2—3	3—4	4—5	5—6	6—7	7—8	8—9	9—10	10—11	11—12
Nombre d'observations	41	34	54	39	49	45	41	33	37	41	47	39

On teste l'hypothèse simple H_1 = {*la position de l'aiguille des heures est uniformément distribuée sur le cadran*} contre l'hypothèse contraire multiple.

Dans cet exemple, $n = 500, p_i = 1/12, i = 1, ..., 12, np_i \approx 41{,}67$. Le théorème 1 nous permet d'admettre approximativement que $\chi^2(X) \in \mathbf{H}_{11}$. Dans cet exemple on s'assure par un calcul immédiat que $\chi^2(X) \approx 10$, et le niveau réel du test du χ^2 est environ égal à $1 - \mathbf{H}_{11}(]10, \infty[) \approx 0{,}47$ (cf. tableau III). Ceci exprime que les résultats de l'expérience s'accordent avec l'hypothèse H_1 du point de vue d'un test du χ^2 de niveau $1 - \epsilon$ compris entre 0,47 et 1.

Nous avons déjà mentionné que le test du χ^2 est largement répandu. Son champ d'application ne se limite pas aux seules hypothèses simples. Nous nous en assurerons dans le paragraphe suivant.

*) Cet exemple a été emprunté à [19].

§ 17. Test d'hypothèses relatives à l'appartenance de la loi de l'échantillon à une famille paramétrique

Considérons le problème de choix entre l'hypothèse multiple $H_1 = \{X \in \{\mathbf{P}_\alpha\}_{\alpha \in A}\}$ et son alternative $H_2 = \{X \in \mathbf{P}, \mathbf{P} \notin \{\mathbf{P}_\alpha\}_{\alpha \in A}\}$. L'hypothèse H_1 peut par exemple consister à vérifier que X est tiré d'une population normale.

Un autre exemple d'hypothèse H_1 est que $X \in \mathbf{B}_{\theta(\alpha)}$, où $\dim \alpha < \dim \theta$. Ce problème peut bien sûr être traité comme un problème de test de l'hypothèse que X est distribué suivant une loi d'une sous-famille paramétrique (cf. § 15) mais la première interprétation est exacte aussi, car dans le cas où l'expérience ne donne lieu qu'à un nombre fini d'issues possibles (cf. définition de $\mathbf{B}_\theta$ dans le § 2.2), la famille $\{\mathbf{B}_\theta\}$ contient *toutes* les distributions possibles de l'échantillon.

Dans le numéro suivant on étudiera le problème de test de l'hypothèse $\{X \in \mathbf{B}_{\theta(\alpha)}\}$ et on montrera que le problème général d'appartenance de la loi de l'échantillon à une famille paramétrique peut être ramené au premier par un groupement des données.

1. Test de l'hypothèse $\{X \in \mathbf{B}_{\theta(\alpha)}\}$. Groupement des données. Considérons le problème général formulé au début du paragraphe dans le cas d'un espace arbitraire $\mathscr{X}$. Prenons une partition de $\mathscr{X}$ en domaines (« intervalles ») $\{\Delta_1, \ldots, \Delta_r\}$ telle que le nombre r d'« intervalles » soit supérieur à $l + 1$, où $l = \dim \alpha$. Groupons les observations sur ces intervalles. Si l'hypothèse $H_1 = \{X \in \mathbf{P}_\alpha\}$ est vraie, les probabilités que les observations tombent dans les intervalles Δ_i seront égales à

$$p_i(\alpha) = \mathbf{P}_\alpha(\Delta_i).$$

Ceci exprime que dans ce cas le vecteur $\theta = (\theta_1, \ldots, \theta_r)$ des probabilités que les observations tombent dans Δ_i doit être porté par la courbe $\theta = p(\alpha) = (p_1(\alpha), \ldots, p_r(\alpha))$.

Nous devons donc, au vu de l'échantillon $Y \in \mathbf{B}_\theta$ obtenu par groupement, vérifier l'hypothèse H_1 que Y est distribué suivant une loi de la sous-famille paramétrique $\{\mathbf{B}_{p(\alpha)}\}$, contre l'hypothèse alternative $\{Y \in \mathbf{B}_\theta\}$, où θ n'est pas situé sur la courbe $\theta = p(\alpha)$, $\alpha \in A$. Nous avons déjà envisagé ce problème au § 15 où nous avons trouvé un test asymptotiquement minimax de choix entre H_1 et l'alternative voisine

$$H_2 = \{Y \in \mathbf{B}_\theta, \inf_\gamma |\theta - p(\alpha_0 + \gamma n^{-1/2}) I^{1/2}(p(\alpha_0 + \gamma n^{-1/2}))| > bn^{-1/2}\} \qquad (1)$$

(cf. remarque 15.3 suivant le théorème 15.4. Le point α_0 désigne une valeur « localisée » du paramètre telle que les alternatives soient situées dans un voisinage du point $\theta_0 = p(\alpha_0)$.) Le test du rapport de vraisemblance (15.11)

devient ici

$$\ln R_1(X) = \max_\theta \sum_{i=1}^r \nu_i \ln \theta_i - \max_\alpha \sum \nu_i \ln p_i(\alpha) > h_\epsilon/2,$$

ou ce qui est équivalent

$$\sum_{i=1}^r \nu_i \ln \frac{\nu_i}{np_i(\hat{\alpha}^*)} > h_\epsilon/2,$$

où $\hat{\alpha}^*$ est un estimateur du maximum de vraisemblance de α au vu de Y (ou de $\nu = (\nu_1, \ldots, \nu_r)$). Ce test est asymptotiquement équivalent (cf. théorème 15.4) au test

$$(p(\hat{\alpha}^*) - \nu n^{-1}) I(p(\hat{\alpha}^*))(p(\hat{\alpha}^*) - \nu n^{-1})^T > h_\epsilon.$$

Vu que l'on connaît la forme de la matrice $I(\theta)$ (cf. (16.5)), en se servant de (16.9) on déduit du théorème 15.4 le

COROLLAIRE 1. *Si $r - 1 > l$ et la fonction $p(\alpha)$ vérifie les conditions du théorème* 15.4, *le test du rapport de vraisemblance de niveau asymptotique $1 - \epsilon$ de l'hypothèse $H_1 = \{X \in \mathbf{P}_\alpha, \mathbf{P}_\alpha \in \{\mathbf{P}_\alpha\}_{\alpha \in A}\}$ contre l'hypothèse complémentaire H_2 d'après les données groupées est asymptotiquement minimax* (*de H_1 contre* (1)) *et est de la forme*

$$\sum_{i=1}^r \nu_i \ln \frac{\nu_i}{np_i(\hat{\alpha}^*)} > h_\epsilon/2, \tag{2}$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ d'une distribution du χ^2 à $r - l - 1$ degrés de liberté. Ce test est asymptotiquement équivalent au test

$$\hat{\chi}^2(X) = \sum_{i=1}^r \frac{(\nu_i - np_i(\hat{\alpha}_i^*))^2}{np_i(\hat{\alpha}_i^*)} > h_\epsilon. \tag{3}$$

Le dernier test s'appelle aussi *test du χ^2* lorsque ce sont les paramètres fantômes inconnus qui sont estimés au vu d'un échantillon. La distribution de la statistique $\hat{\chi}^2(X)$ converge pour l'hypothèse H_1, ainsi qu'il ressort du corollaire 1, vers une distribution du χ^2 à $r - l - 1$ degrés de liberté (le nombre $r - 1$ de degrés de liberté dans la distribution limite de la statistique $\chi^2(X)$ a baissé du nombre de paramètres scalaires $\alpha_1, \ldots, \alpha_l$ estimés au vu de l'échantillon X).

EXEMPLE 1. Dans l'exemple 2.26.3 on a décrit le mécanisme de transmission des groupes sanguins 0, A, B, AB. Ce mécanisme est géré par des gènes de type A, B et 0. Désignons par p, q et $r = 1 - p - q$ les probabilités d'apparition de ces gènes dans une population donnée. Les probabilités $p_i(\alpha)$ qu'un individu soit du groupe i ont été déterminées dans l'exemple 2.26.3 et rassemblées dans le tableau 1 du § 26.

On dispose d'un échantillon X de fréquences ν_i, $i = 1, 2, 3, 4$, (cf. tableau 1) d'apparition du groupe i, obtenu par un sondage de $n = 353$ personnes. Dans l'exemple 2.26.3 on a trouvé pour cet échantillon les valeurs approchées de l'estimation du maximum de vraisemblance $\hat{\alpha}^* = (p^*, q^*) = (0{,}246, 0{,}173)$. Ceci nous donne les valeurs $p_i(\hat{\alpha}^*)$ du tableau 1.

Tableau 1. **Répartition des personnes sondées d'après leur groupe sanguin**

	0	A	B	AB	Total
ν_i	121	120	79	33	353
$p_i^* = \nu_i n^{-1}$	0,343	0,340	0,224	0,093	1
$p_i(\hat{\alpha}^*)$	0,337	0,347	0,231	0,085	1

Nous obtenons la possibilité d'appliquer le corollaire 1 pour vérifier l'hypothèse que le mécanisme de transmission du groupe sanguin se déroule bien tel qu'on l'a décrit. En se servant des données du tableau, on trouve que la statistique $\hat{\chi}^2(X)$ (cf. (3)) est égale ici à environ 0,44. Ce résultat s'accorde bien avec l'hypothèse, puisque la valeur critique h_ϵ correspondant à la distribution du χ^2 à un degré de liberté et à la valeur $\epsilon = 0{,}2$ est égale à $h_{0,2} \approx 1{,}64$.

EXEMPLE 2. *Problème des caractères contingents.* Supposons qu'un échantillon X est le résultat d'un sondage d'objets dont on a mesuré les caractères A et B. Le premier est susceptible de prendre les valeurs $A_1, \ldots$ $\ldots, A_s$, le second, les valeurs $B_1, \ldots, B_t$. On demande si ces caractères sont dépendants entre eux. Nous pouvons par exemple effectuer une expérience G d'issues $B_1, \ldots, B_t$ dans des conditions différentes $A_1, \ldots, A_s$. Le problème consiste à dire si les résultats de l'expérience G dépendent des conditions de leur réalisation.

Ce problème peut être traité aussi comme un problème de test de l'indépendance de deux variables aléatoires ξ et η au vu d'observations groupées sur le couple (ξ, η).

Les résultats des expériences se présentent ici sous la forme d'une matrice $\|\nu_{ij}\|$, où ν_{ij} est le nombre d'apparitions des issues A_i et B_j dans un échantillon X de taille n (chaque élément de cet échantillon est un couple de caractères de l'objet étudié).

Posons $p_{ij} = \mathbf{P}(A_i B_j)$, $p_{i\cdot} = \sum_{j=1}^{t} p_{ij}$, $p_{\cdot j} = \sum_{i=1}^{s} p_{ij}$. L'hypothèse H_1 d'indépendance des caractères sera alors de la forme $H_1 = \{p_{ij} = p_{i\cdot} p_{\cdot j}\}$. Il est immédiat de voir que cette hypothèse concerne l'appartenance de la distribution de l'échantillon à une sous-famille paramétrique, où le rôle du paramètre α est tenu par un vecteur $(s + t - 2)$-dimensionnel $\alpha = (p_{1\cdot}, \ldots, p_{s-1\cdot}, p_{\cdot 1}, \ldots, p_{\cdot t-1})$ (les valeurs $p_{s\cdot}$ et $p_{\cdot t}$ sont tirées des égalités $p_{s\cdot} = 1 - \sum_{i=1}^{s-1} p_{i\cdot}$, $p_{\cdot t} = 1 - \sum_{j=1}^{t-1} p_{\cdot j}$).

La fonction de vraisemblance de l'échantillon X pour H_1 est égale à

$$\prod_{i,j} p_{ij}^{\nu_{ij}} = \prod_{i} p_{i\cdot}^{\nu_{i\cdot}} \prod_{j} p_{\cdot j}^{\nu_{\cdot j}}, \quad \nu_{i\cdot} = \sum_{j=1}^{t} \nu_{ij}, \quad \nu_{\cdot j} = \sum_{i=1}^{s} \nu_{ij}.$$

Des résultats du § 16 (comparer avec (16.1)) il découle que l'estimateur du maximum de vraisemblance $\hat{\alpha}^*$ pour une telle fonction est

$$p_{i\cdot}^* = \nu_{i\cdot}/n, \quad \hat{p}_{\cdot j}^* = \nu_{\cdot j}/n.$$

Le test du χ^2 est donc ici de la forme

$$\hat{\chi}^2(X) = \sum_{i,j} \frac{(\nu_{ij} - n\hat{p}_{i\cdot}^* \hat{p}_{\cdot j}^*)^2}{n\hat{p}_{i\cdot}^* \hat{p}_{\cdot j}^*} = n \sum_{i,j} \frac{(\nu_{ij} - n^{-1}\nu_{i\cdot}\nu_{\cdot j})^2}{\nu_{i\cdot}\nu_{\cdot j}} > h_\epsilon,$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à $st - 1 - (s + t - 2) = (s - 1)(t - 1)$ degrés de liberté.

Tableau 2

A \ B	0—1	1—2	2—3	≥3	$\nu_{i\cdot}$
0	2161	3577	2184	1636	9558
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	225	419	96	38	778
≥4	39	98	31	14	182
$\nu_{\cdot j}$	6116	10928	5173	3016	25263

On pourrait citer une foule de problèmes d'application faisant intervenir le test des caractères contingents construit ci-dessus. Considérons à titre d'illustration un problème de sondage sociologique portant sur le lien entre le budget d'une famille et le nombre d'enfants de cette famille (cf. [19]).

EXEMPLE 2A. Supposons que le caractère A désigne le nombre d'enfants et prend les valeurs 0, 1, 2, 3, $\geqslant 4$. Le caractère B indique l'une des fourchettes $(0 - 1)$, $(1 - 2)$, $(2 - 3)$, $(\geqslant 3)$ du budget (une unité représente 1000 couronnes suédoises). Les résultats d'un sondage portant sur $n = 25\,263$ familles ont été rassemblés dans le tableau 2.

Dans cet exemple, $\hat{\chi}^2(X) = 568{,}5$, quantité qui est considérablement plus grande que la valeur critique h_ϵ de la distribution du χ^2 à $(5 - 1)(4 - - 1) = 12$ degrés de liberté même pour les ϵ assez petits. Force est donc d'infirmer l'hypothèse $H_1 = \{A$ *et* B *sont indépendants* (*non contingents*)$\}$.

A noter toutefois qu'une analyse plus fine met en évidence la très faible dépendance des caractères A et B.

2. Cas général. Le test du χ^2 de ce problème possède les mêmes défauts que dans les problèmes du paragraphe précédent.

Le problème de test de l'hypothèse $\{X \in \mathbf{P}_\theta\}$ que la loi de X appartient à une famille paramétrique $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ admet bien sûr une approche plus large identique à celle qui a été exposée au § 12. Définissons une distance $d(\mathbf{P}, \mathbf{Q})$ sur l'espace des distributions. Trouvons ensuite le point $\mathbf{P}_{\theta^*}$ de $\{\mathbf{P}_\theta\}$ le plus proche de $\mathbf{P}_n^*$ pour la distance d. Pour $\mathbf{P}_{\theta^*}$ on peut prendre également $\mathbf{P}_{\hat{\theta}^*}$ où $\hat{\theta}^*$ est un estimateur du maximum de vraisemblance (cf. § 2.5) ou un autre estimateur raisonnable. La distance $d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*)$ sera petite ou grande selon que l'hypothèse H_1 ou l'hypothèse H_2 sera vraie. Ceci nous suggère la recette suivante du test : l'hypothèse H_1 est rejetée si $d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*) > c$ et acceptée dans le cas contraire.

Le nombre c doit être choisi tel que

$$\sup_{\theta\in\Theta} \mathbf{P}_\theta(d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*) > c) \leqslant \epsilon,$$

ou tel que cette relation soit réalisée asymptotiquement. Le corollaire 1 nous suggère de prendre pour distance $d(\mathbf{P}_{\theta^*}, \mathbf{P}_n^*)$ les statistiques de (2) et (3), statistiques qui, entre autres, présentent encore l'avantage d'être asymptotiquement non paramétriques : la distribution limite du $\hat{\chi}^2(X)$ par exemple ne dépend pas de θ pour l'hypothèse $H_1 = \{X \in \mathbf{P}_\theta\}$.

Voyons comment l'approche générale développée ci-dessus se réalise dans deux cas particuliers importants où les familles paramétriques dépendent des paramètres de translation et d'échelle.

1) Soit à tester l'hypothèse $X \in \mathbf{P}_\theta$, $\theta \in R$, où $\mathbf{P}_\theta(A) = \mathbf{P}(A - \theta)$, $A \subset R$. Désignons par $F(x)$ la fonction de répartition de $\mathbf{P}$ et posons $F_\theta(x) = F(x - \theta)$. Pour d nous prendrons la distance utilisée dans le test de Kolmogorov.

THÉORÈME 1. *Supposons que* $X \in \mathbf{P}_\theta$, $F_\theta(x) = F(x - \theta)$ *et que la fonction* $F(x)$ *admet une densité de probabilité bornée uniformément continue* $f(x) = F'(x)$, $\int x^2 f(x)dx < \infty$. *Si*

l'on désigne $\int xf(x)dx = a, \theta^* = \bar{x} - a$, *on a pour tout* θ

$$\lim_{n\to\infty} \mathbf{P}_\theta(\sup_x \sqrt{n}\,|F_n^*(x) - F_{\theta^*}(x)| > c) =$$

$$= \mathbf{P}\Big(\sup_x \Big| w^\circ(F(x)) + f(x) \int w^\circ(F(t))dt \Big| > c\Big),$$

où w° *est un pont brownien.*

Le second membre de cette relation est indépendant de θ. En le calculant pour F donnée et en choisissant $c = c_\epsilon$ de telle sorte qu'il soit égal à ϵ, on obtient le test

$$D_n \equiv \sup_x \sqrt{n}\,|F_n^*(x) - F(x - \theta^*)| > c_\epsilon$$

de niveau asymptotique $1 - \epsilon$, relatif à l'hypothèse H_1 que la loi de X appartient à la famille paramétrique $\{\mathbf{P}_\theta\}$, où θ est un paramètre de translation.

DÉMONSTRATION du théorème 1. Considérons le processus

$$W_n(x) = \sqrt{n}(F_n^*(x) - F_{\theta^*}(x)) = w_n(x) - \sqrt{n}(F_{\theta^*}(x) - F_\theta(x)),$$

où $w_n(x) = \sqrt{n}(F_n^*(x) - F_\theta(x))$. Pour $t \to \theta$, on a

$$F_t(x) - F_\theta(x) = -(t - \theta)(f(x - \theta) + \epsilon(t, \theta, x)),$$

$$|\epsilon(t, \theta, x)| \leqslant \omega_{|t-\theta|},$$

où ω_Δ, le module de continuité de f, est indépendant de x, $\omega_\Delta \to 0$ pour $\Delta \to 0$. Puisque $\theta^* \underset{P_\theta}{\to} \theta$, en admettant que $t = \theta^*$ et en posant, sans nuire à la généralité, $a = 0$, on obtient

$$\sqrt{n}(F_{\theta^*}(x) - F_\theta(x)) = -f(x - \theta) \int td[\sqrt{n}(F_n^*(t) - F_\theta(t))] + \epsilon(\theta^*, \theta, x) =$$

$$= -f(x - \theta) \int tdw_n(t) + \epsilon(\theta^*, \theta, x),$$

$$|\epsilon(\theta^*, \theta, x)| \leqslant \omega(\theta^* - \theta) \equiv \sqrt{n}\,|\theta^* - \theta|\,\omega_{|\theta^*-\theta|} \underset{P_\theta}{\to} 0.$$

Pour tout $N > 0$, la fonctionnelle

$$H_N(w_n) = \sup_x \Big| w_n(x) - f(x - \theta) \int_{-N}^{N} w_n(t)dt \Big|$$

est continue pour une métrique uniforme. Cette propriété est préservée par le changement de x en $F_\theta^{-1}(y) = \theta + F^{-1}(y)$ qui est nécessaire à l'application du théorème 1.6.3. En vertu de ce dernier on a

$$H_N(w_n) \Rightarrow \sup_x \Big| w^\circ(F(x - \theta)) + f(x - \theta) \int_{-N}^{N} w^\circ(F(t - \theta))dt \Big|.$$

Pour établir la relation annoncée

$$D_n \Rightarrow \sup_x |w^\circ(F(x - \theta)) + f(x - \theta) \int w^\circ(F(t - \theta))dt|$$

(la θ-translation ne modifie pas la valeur du second membre), il nous reste, en vertu des relations

$$|D_n - H_N(w_n)| \leqslant \omega(\theta^* - \theta) + c \int_{|t| \geqslant N} w_n(t)dt, \tag{4}$$

$$\omega(\theta^* - \theta) \underset{P_\theta}{\to} 0,$$

à nous assurer que l'intégrale de (4) et l'intégrale $\int_{|t| \geqslant N} w^\circ(F(t))dt$ (pour simplifier nous posons $\theta = 0$) convergent en probabilité vers 0 lorsque $n \to \infty$ et $N \to \infty$. La meilleure façon d'estimer ces deux intégrales est, de toute évidence, de prouver que leurs variances sont petites et d'utiliser l'inégalité de Tchébychev. Vu que les moments du premier et du deuxième ordre des intégrants de ces deux intégrales se comportent de la même façon, nous pouvons nous contenter d'estimer l'une d'elles, par exemple

$$\int_{-\infty}^{-N} w^\circ(F(t))dt.$$

La relation $\mathbf{E}w^\circ(s)w^\circ(u) = \min(s, u) + su \leqslant 2 \min(s, u)$ pour $s \leqslant 1$ et $u \leqslant 1$ nous donne

$$\mathbf{E}\left(\int_{-\infty}^{-N} w^\circ(F(t))dt\right)^2 \leqslant 2 \int_{-\infty}^{-N}\int_{-\infty}^{-N} \min(F(t), F(s))dtds =$$

$$= 4 \int_{-\infty}^{-N} (-t - N)F(t)dt \leqslant -8 \int_{-\infty}^{-N} tF(t)dt \to 0$$

lorsque $N \to \infty$, puisque $\int t^2 dF(t) < \infty$. Les autres intégrales se traitent de façon analogue. ◄

2) Soit à tester maintenant l'hypothèse $X \in \mathbf{P}_\theta$, $\theta \in R$, $\theta > 0$, où $\mathbf{P}_\theta(A) = \mathbf{P}(A|\theta)$, $A \subset R$. Désignons encore par F la fonction de répartition de $\mathbf{P}$ et posons

$$\sigma^2 = \mathbf{E}_1 x_1^2 = \int x^2 \mathbf{P}(dx), \quad \theta^* = \frac{1}{\sigma}\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}.$$

THÉORÈME 2. *Supposons que $X \in \mathbf{P}_\theta$, $F_\theta(x) = F(x/\theta)$ et qu'il existe une densité continue bornée $f(x) = F'(x)$ telle que*

$$\sup_x |xf(x)| < \infty, \quad \int x^4 f(x)dx < \infty. \tag{5}$$

Alors pour tout θ

$$\lim_{n\to\infty} \mathbf{P}_\theta\left(\sup_x \sqrt{n}\,|F_n^*(x) - F(x/\theta^*)| > c\right) =$$

$$= \mathbf{P}\left(\sup_x \left| w^\circ(F(x)) + xf(x) \int tw^\circ(F(t))dt \right| > c\right).$$

DÉMONSTRATION. Elle reprend *ad litteram* celle du théorème 1. On a

$$W_n(x) = \sqrt{n}(F_n^*(x) - F(x/\theta^*)) = w_n(x) - \sqrt{n}(F(x/\theta^*) - F(x/\theta)),$$
$$w_n(x) = \sqrt{n}(F_n^*(x) - F(x/\theta)).$$

Pour $t \to \theta$,

$$F_t(x) - F_\theta(x) = \left(\frac{x}{t} - \frac{x}{\theta}\right)\left(f\left(\frac{x}{\theta}\right) + \varepsilon(t, \theta, x)\right),$$

où, en vertu de la relation $f(x) < c/|x|$ et de la continuité uniforme de f, on a sur tout intervalle fini $\sup_x |\varepsilon(t, \theta, x)| \leqslant \omega_{|t-\theta|} \to 0$. En admettant que $t = \theta^* \underset{P_\theta}{\to} \theta$, on obtient

$$\sqrt{n}\left(F\left(\frac{x}{\theta^*}\right) - F\left(\frac{x}{\theta}\right)\right) =$$
$$= \sqrt{n}\left(\frac{x}{\theta^*} - \frac{x}{\theta}\right)f\left(\frac{x}{\theta}\right) - \frac{\sqrt{n}(\theta^* - \theta)}{\theta^*}\cdot\frac{x}{\theta}\cdot f\left(\frac{x}{\theta}\right)\varepsilon(\theta^*, \theta, x),$$

où $\sup_x$ du second terme converge en $\mathbf{P}_\theta$-probabilité vers 0. Reste à appliquer les raisonnements du théorème précédent (la petitesse des intégrales $\int_{|t|>N} tw^\circ(F(t))dt$ et $\int_{|t|>N} tw_n(t)dt$ est assurée par la condition (5)) et à remarquer que la partie principale de $W_n(x)$ est égale à (on convient sans nuire à la généralité que $\sigma^2 = 1$)

$$w_n(x) - \frac{\sqrt{n}x(\theta^{*2} - \theta^2)}{\theta\theta^*(\theta + \theta^*)}f(x/\theta) = w_n(x) - \frac{xf(x/\theta)}{\theta\theta^*(\theta + \theta^*)}\int t^2 dw_n(t) =$$
$$= w_n(x) - \frac{2xf(x/\theta)}{\theta\theta^*(\theta + \theta^*)}\int tw_n(t)dt,$$

où $\theta^*(\theta + \theta^*) \underset{P_\theta}{\to} 2\theta^2$. Donc,

$$\sup_x |W_n(x)| \to \sup_x \left| w^\circ\left(F\left(\frac{x}{\theta}\right)\right) + \theta^{-3}xf\left(\frac{x}{\theta}\right)\int tw^\circ\left(F\left(\frac{t}{\theta}\right)\right)dt\right| =$$
$$= \sup_x \left| w^\circ\left(F\left(\frac{x}{\theta}\right)\right) + \frac{x}{\theta}f\left(\frac{x}{\theta}\right)\int tw^\circ(F(t))dt\right|.$$

Ce qui prouve le théorème 2, puisque la contraction effectuée sur x sous le signe $\sup_x$ est sans effet.

Le lecteur peut établir des résultats identiques pour les statistiques $\int (F_n^*(x) - F_{\theta^*}(x))^2 dF_{\theta^*}(x)$.

§ 18. Stabilité des décisions statistiques

Dans les problèmes d'estimation ou de test d'hypothèses envisagés dans les paragraphes précédents, nous avons posé à chaque fois un certain nombre de conditions en construisant des procédures statistiques. Ces conditions portaient en particulier sur l'indépendance des observations et sur

leur équidistribution, ainsi que sur le caractère de la distribution **P** des éléments de l'échantillon. La non-réalisation de ces conditions aurait mis en défaut les conclusions respectives (relatives par exemple au caractère de la distribution limite ou à l'optimalité de telle ou telle statistique).

D'autre part, les conditions discutées en pratique sont en règle générale le résultat d'une approximation et d'une inévitable idéalisation. Donc, ces conditions ne sont pas remplies exactement, d'où la crainte que les recommandations prodiguées à l'aide de telle ou telle procédure statistique ne soient pas fondées.

Par conséquent, comme dans tout domaine des mathématiques lié aux applications, il est nécessaire, avant la mise en œuvre de ces méthodes, de fixer la marge des écarts par rapport aux hypothèses admises pour remettre éventuellement en cause les résultats obtenus.

Du point de vue mathématique, ce problème est très voisin du problème de stabilité *).

Les écarts les plus courants par rapport aux conditions mentionnées sont de la nature suivante.

1) La série d'observations X contient un faible pourcentage de valeurs aberrantes, c'est-à-dire des observations entachées de grossières erreurs de mesure ou d'enregistrement, ou engendrées par un autre mécanisme « perturbateur » différent du système étudié. Comme il est pratiquement impossible de différencier ces observations des autres, on cherche des procédures peu sensibles à ces « pollutions ».

2) La distribution de x_i n'est égale à **P** qu'approximativement.

3) Les éléments de X ne sont pas indépendants, mais faiblement dépendants.

Le problème consiste à construire, pour les principaux problèmes de statistique, des décisions qui soient par leur efficacité proches des décisions optimales et qui dans le même temps soient insensibles aux écarts par rapport aux hypothèses admises ou, à la rigueur, par rapport à celles qui sont essentielles pour nous. Ce problème qui est très compliqué et pas toujours exactement posé ne peut être considéré comme étudié à fond. Les résultats obtenus étant encore éparses, on ne s'arrêtera que sur quelques exemples typiques.

1. Estimation de la moyenne pour des distributions symétriques. Soit $X \in \mathbf{P}$, où **P**, distribution sur une droite, admet la densité $f(t - \alpha)$ par rapport à la mesure de Lebesgue, $f(t) = f(-t)$. Nous étudions les deux estimateurs suivants du paramètre $\alpha = \mathsf{E}x_1$: l'estimateur

$$\alpha^* = \overline{x}$$

*) On se sert aussi du terme de « robustesse ».

et l'estimateur α^{**} basé sur les quantiles empiriques :

$$\alpha^{**} = \frac{1}{r-1} \sum_{k=1}^{r-1} \zeta^*_{kp}, \tag{1}$$

où $0 < p < 1, r = 1/p$ étant un entier. Pour $p = 1/2$, l'estimateur α^{**} se transforme en la médiane empirique $\zeta^* = \zeta^*_{1/2}$.

Bornons-nous pour l'instant au cas $p = 1/2$. Pour $n \to \infty$, on a

$$(\alpha^* - \alpha)\sqrt{n} \in \Phi_{0,\sigma_1^2}, \quad \sigma_1^2 = \int t^2 f(t)dt. \tag{2}$$

Par ailleurs, dans le corollaire 2.2.1 on a vu que pour $n \to \infty$

$$(\alpha^{**} - \alpha)\sqrt{n} \in \Phi_{0,\sigma_2^2}, \quad \sigma_2^2 = \frac{1}{4f^2(0)}. \tag{3}$$

En reprenant la démonstration de ce corollaire, on établit aisément que le terme $x_{(k)}$ de l'échantillon ordonné associé à X aura, pour toute valeur fixe de la différence $k - k_0$, la même distribution limite que $\alpha^{**} = \zeta^* = x_{(k_0)}, k_0 = [(n+1)/2]$.

On en déduit que l'estimateur $\alpha^{**} = \zeta^*$ est insensible (du point de vue de ses propriétés asymptotiques) à l'adjonction à l'échantillon X d'un nombre fini quelconque d'éléments aberrants. En effet, si l'échantillon X contient l éléments aberrants, l'estimation α^{**} sera située entre les valeurs $y_{(k_1)}$ et $y_{(k_2)}$, où $k_1 = k_0 - l, k_2 = k_0 + l$, et $y_{(k)}, k = 1, \ldots, n - l$, est l'échantillon ordonné associé à un échantillon $Y \in \mathbf{P}$ de taille $n - l$. Mais les propriétés asymptotiques de $y_{(k_1)}$ et $y_{(k_2)}$ sont identiques et sont confondues avec celles de la médiane empirique.

Donc, l'estimateur α^{**} est insensible aux aberrations quelles qu'elles soient. On ne peut en dire autant de l'estimateur $\alpha^* = \bar{x}$, où la contribution des aberrations est importante (par exemple si leur nombre l est de l'ordre de n). Il est aisé de comprendre que α^{**} reste stable si le nombre l d'éléments aberrants n'est pas élevé en regard de n. Il le reste encore si l'on remplace ζ^* par une statistique (1) de forme plus générale.

D'autre part, dans le cas particulier important où $\mathbf{P} = \Phi_{\alpha,\sigma_1^2}$, la valeur $\sigma_2^2 = \sigma_1^2\pi/2$ ($f(0) = (\sigma_1\sqrt{2\pi})^{-1}$) est de $\pi/2$ fois plus grande que la variance σ_1^2 de l'estimateur efficace $\alpha^* = \bar{x}$. La différence entre les efficacités de α^{**} et α^* peut être réduite davantage si les estimateurs (1) sont envisagés pour $r = 3, 4$, etc. Nous obtenons alors un estimateur α^{**} presque aussi efficace que $\bar{x}$ (en l'absence d'aberrations), qui en même temps sera stable par rap-

port aux aberrations. Outre (1) on peut prendre la moyenne tronquée

$$\alpha^{**} = \frac{1}{n - 2np} \sum_{k=np+1}^{n-np} x_{(k)}, \tag{4}$$

dont la variance se rapprochera pour les petits p de la variance σ_1^2 de l'estimateur α^*.

Signalons par ailleurs que les propriétés de l'estimateur $\alpha^* = \bar{x}$ dépendent peu des variations de $\mathbf{P}$ n'affectant pas la variance $\sigma_1^2 = \int t^2 f(t)dt$, et notamment des variations locales de $f(t)$ en $t = 0$. En ce sens il est stable. Mais *la propriété d'optimalité* de cet estimateur qui a lieu pour $\mathbf{P} = \Phi_{\alpha, \sigma_1^2}$ *est instable*. En effet, supposons que pour $\epsilon > 0$ petit,

$$\mathbf{P} = (1 - \epsilon)\Phi_{\alpha, 1} + \epsilon \mathbf{U}_{\alpha-\epsilon, \alpha+\epsilon}.$$

Alors $f(0) = (1 - \epsilon)/\sqrt{2\pi} + 1/2 > 1/2$ et comme le montrent les relations (2) et (3), l'estimateur $\alpha^{**} = \zeta^*$ sera sensiblement meilleur (ϵ doit être petit mais pas inférieur à $1/\sqrt{n}$).

D'autre part, l'estimateur $\alpha^{**} = \zeta^*$ (plus exactement sa distribution) est stable par rapport aux variations de $\mathbf{P}$ n'affectant pas la valeur $f(0)$.

Ces remarques peuvent être reformulées sans peine pour les tests, par exemple pour les tests sans biais uniformément les plus puissants $|\bar{x} - \alpha_0| > c$ de l'hypothèse $H_1 = \{\alpha = \alpha_0\}$ contre $H_2 = \{|\alpha - \alpha_0| > d > 0\}$ au vu d'un échantillon $X \in \Phi_{\alpha, 1}$.

2. Statistiques t et S_0^2. Considérons maintenant le problème de la stabilité des procédures statistiques (estimation et test d'hypothèses) faisant intervenir les statistiques

$$t = \frac{(\bar{x} - \alpha)\sqrt{n}}{S_0}, \quad S_0^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

On sait que ces statistiques (cf. §§ 3.7, 3.8) sont à la base de tests optimaux de choix entre hypothèses relatives à la moyenne α et à la variance σ^2 de populations normales dans le cas où le second paramètre (σ^2 ou α) de la distribution Φ_{α, σ^2} est inconnu.

Les statistiques t et S_0^2 se conduisent de manière différente face aux violations de la condition $X \in \Phi_{\alpha, \sigma^2}$. Supposons que n est grand et que $X \in \mathbf{P}$, où $\mathbf{P}$ est une distribution quelconque de moyenne α et de variance finie. La distribution de t se laisse approcher, comme pour le cas où $X \in \Phi_{\alpha, \sigma^2}$, par la loi normale réduite $\Phi_{0, 1}$. Ceci résulte des théorèmes de con-

tinuité (§ 1.5) et du fait que

$$(\bar{x} - \alpha)\sqrt{n}/\sqrt{\mathbf{V}x_1} \in \Phi_{0,1}, \qquad S_0^2 \underset{P}{\to} \mathbf{V}x_1.$$

Ce qui vient d'être dit exprime que le niveau du test de Student sera peu différent du niveau donné pour les grands n, même si la distribution $\mathbf{P}$ de l'échantillon X s'éloigne considérablement de la distribution normale.

On ne peut en dire autant des tests construits à l'aide de la statistique S_0^2. Cette circonstance est liée au fait que la distribution limite de S_0^2 dépend de la valeur $\mathbf{E}x_i^4$. En effet, des considérations du chapitre 1, il résulte que

$$(S_0^2 - \sigma^2)\sqrt{n} \in \Phi_{0,d^2}, \qquad d^2 = \mathbf{E}(x_1^2 - \sigma^2)^2 = \mathbf{V}x_1^2.$$

Donc, le niveau du test construit à l'aide de la statistique S_0^2 pour une population normale peut différer considérablement du niveau donné si $X \in \mathbf{P}$ et $\mathbf{P} \neq \Phi_{\alpha,\sigma^2}$ (ces niveaux seront égaux si les moments d'ordre 4 de $\mathbf{P}$ et de Φ_{α,σ^2} sont confondus).

Les statistiques t et S_0^2 sont sensibles au refus de l'hypothèse d'indépendance des observations de l'échantillon X. Si par exemple les observations sont corrélées et que le coefficient de corrélation soit égal à ρ, en convenant sans nuire à la généralité que $\alpha = 0$, on obtient

$$\mathbf{E}S_0^2 = \frac{1}{n-1}\mathbf{E}\left[\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2\right] = \frac{1}{n-1}\left[n\sigma^2 - \frac{1}{n}\mathbf{E}\left(\sum_{i=1}^{n} x_i\right)^2\right] =$$

$$= \frac{1}{n-1}[n\sigma^2 - \sigma^2(1-\rho) - n\sigma^2\rho] = \sigma^2(1-\rho).$$

Donc, même la propriété d'absence de biais de S_0^2 est violée, bien que l'écart ne soit pas élevé pour les petits ρ. La détermination des distributions de t et de S_0^2 lorsque les observations sont dépendantes est un problème très compliqué.

3. Test du rapport de vraisemblance. Ce test est en principe très sensible à la présence des aberrations et même aux petits écarts par rapport aux hypothèses relatives à la distribution de X. Supposons par exemple que l'on teste les deux hypothèses simples $H_1 = \{X \in \Phi_{0,1}\}$ et $H_2 = \{X \in \mathbf{U}_{-1,1}\}$. Il est évident que si l'on se sert d'un test le plus puissant de Neyman-Pearson, l'apparition d'au moins une observation x à l'extérieur de l'intervalle $[-1, 1]$, les autres observations étant idéalement distribuées suivant la loi uniforme $\mathbf{U}_{-1,1}$, nous contraindra (avec un risque nul !) d'accepter l'hypothèse H_1. Ceci exprime que la présence d'au moins une aberration ou l'apparition d'écarts, même petits, par rapport à la distribution $\mathbf{U}_{-1,1}$ peuvent nous obliger à prendre une fausse décision.

De ce point de vue, le test de Kolmogorov par exemple est bien plus stable (quoique moins puissant pour H_2). D'une façon générale, les tests non paramétriques sont, comme il faut s'y attendre, bien plus stables que les tests « individuels » optimaux dans tel ou tel problème concret.

S'agissant du problème de décision entre l'hypothèse H_1 que X est normal et l'hypothèse H_2 qu'il est uniforme, on peut chercher des tests puissants et à la fois stables pour les aberrations en se servant comme précédemment du rapport de vraisemblance, mais pour des échantillons « tronqués » (comparer avec (4)). On peut aussi essayer de trouver un autre test. Le choix est riche et souvent il est guidé non seulement par des considérations de stabilité mais aussi par la commodité des calculs.

CHAPITRE 4

PROBLÈMES DE STATISTIQUE À DEUX ÉCHANTILLONS ET PLUS

§§ 1, 2 : problèmes d'homogénéité de deux échantillons.
§ 3 : problèmes de régression.
§ 4 : principaux résultats de l'analyse de variance.
§ 5 : problèmes d'analyse discriminante.

§ 1. Tests d'hypothèses d'homogénéité (totale ou partielle) dans le cas paramétrique

1. Classe de problèmes envisagée. Dans les chapitres précédents, nous avons étudié essentiellement un échantillon X de taille n distribué suivant une loi $\mathbf{P}$ totalement ou partiellement inconnue. Nous passons maintenant aux problèmes faisant intervenir *deux* ou *plus de deux* échantillons.

L'une des classes essentielles de problèmes traités sera celle des problèmes de test d'*homogénéité* (totale ou partielle) de deux échantillons.

Font partie de cette classe les trois types de problèmes suivants :

I. *Test d'homogénéité « ordinaire »*. Le problème consiste ici à éprouver l'hypothèse que deux échantillons X et Y sont distribués suivant la même loi inconnue. Ces problèmes se présentent par exemple lors de la comparaison de deux méthodes de traitement dans un processus technologique ou en agriculture. La comparaison est effectuée généralement par le biais de caractéristiques numériques du produit final (de l'échantillon) qui sont aléatoires. On est confronté à de tels problèmes lorsqu'on teste l'effet d'un nouveau remède en comparant le groupe d'expérience des patients au groupe témoin.

L'exemple 3 de l'Introduction est un problème d'homogénéité.

On étudiera le *cas paramétrique* dans ce paragraphe. Soient donnés une famille paramétrique de distributions $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ et deux échantillons indépendants $X = (x_1, \ldots, x_{n_1})$ et $Y = (y_1, \ldots, y_{n_2})$ de tailles respectives n_1 et n_2 dont on sait *a priori* qu'ils sont distribués suivant une loi de la famille $\{\mathbf{P}_\theta\}$:

$$X \in \mathbf{P}_{\theta_1}, \quad Y \in \mathbf{P}_{\theta_2} \tag{1}$$

pour certains θ_1 et θ_2. Le problème d'homogénéité ordinaire consiste à choisir entre l'hypothèse $H_1 = \{\theta_1 = \theta_2\}$ et son alternative $H_2 = \{\theta_1 \neq \theta_2\}$. Il est évident que les hypothèses H_1 et H_2 sont toutes deux multiples.

II. *Test d'homogénéité en présence d'un paramètre fantôme.* On admet que dim $\theta > 1$. Représentons le vecteur θ sous la forme $\theta = (u, v)$, où u et v sont des « sous-vecteurs » et désignons par u_j et v_j les coordonnées du vecteur θ_j dans (1), $j = 1, 2$.

Supposons que l'on sache que le « sous-paramètre » inconnu v est le même pour les deux échantillons : $v_1 = v_2 = v$. On demande de choisir entre l'hypothèse $H_1 = \{u_1 = u_2\}$ et son alternative $H_2 = \{u_1 \neq u_2\}$.

Ceci est le problème d'homogénéité en présence d'un paramètre fantôme v. Il se distingue du problème d'homogénéité ordinaire par le fait que l'hypothèse contraire de $H_1 = \{\theta_1 = \theta_2\}$ est de la forme $H_2 = \{u_1 \neq u_2, v_1 = v_2\}$.

Les problèmes de cette nature se présentent par exemple dans la situation suivante. Supposons que l'on s'intéresse à l'état d'un objet caractérisé par un vecteur a qui n'est pas mesurable directement mais seulement en présence d'un bruit aléatoire. La nature de ce bruit ne change pas d'une observation à l'autre. Il faut tester l'hypothèse que a est invariant dans deux séries d'observations X et Y.

Si par exemple les mesures effectuées sont de la forme $x_i = a_1 + \xi_i$, où $\xi_i \in \Phi_{\lambda,\sigma^2}$ caractérisent le bruit, et les observations y_i sont de même nature lorsqu'on remplace a_1 par a_2, on peut écrire $X \in \Phi_{a_1+\lambda,\sigma^2}$, $Y \in \Phi_{a_2+\lambda,\sigma^2}$. Nous sommes amenés à considérer le problème de test de l'égalité des moyennes $\{\alpha_1 = \alpha_2\}$ de deux lois normales Φ_{α_1,σ^2} et Φ_{α_2,σ^2} ayant la même variance inconnue σ^2.

III. *Test d'homogénéité partielle.* On teste l'hypothèse H_1 d'une coïncidence « partielle » de θ_1 et θ_2. Plus exactement, on éprouve l'hypothèse $H_1 = \{u_1 = u_2\}$ (les notations sont celles du n° II) contre $H_2 = \{u_1 \neq u_2\}$. Les valeurs v_1 et v_2 peuvent être différentes pour X et Y.

Supposons par exemple que l'on teste en laboratoire le résultat de l'effet d'une nouvelle méthode de traitement sur la productivité d'une céréale. Les observations portent sur le poids total des grains des épis. Supposons que $x_i \in \Phi_{\alpha_1,\sigma_1^2}$, $i = 1, ..., n_1$ pour le lot expérimental et $y_i \in \Phi_{\alpha_2,\sigma_2^2}$ pour le lot de contrôle. Il est naturel d'admettre que la « dispersion » σ^2 varie avec le procédé de traitement. Mais ce qui est essentiel pour nous, c'est de savoir si a varié le principal indice α de productivité. Nous sommes ainsi conduits au problème de décision entre les hypothèses $H_1 = \{\alpha_1 = \alpha_2\}$ et $H_2 = \{\alpha_1 \neq \alpha_2\}$ pour des lois normales dont les variances peuvent être

différentes. Ce problème est bien connu sous le nom de *problème de Berens-Fisher**).

Dans ce paragraphe, nous ramènerons les trois types de problèmes pour des familles paramétriques quelconques au problème, étudié au § 3.15, d'appartenance de la loi d'un échantillon à une sous-famille paramétrique et trouverons la forme des tests asymptotiquement minimax sous la condition que les hypothèses testées soient proches. Ce seront des tests du rapport de vraisemblance qui, pour les lois normales, seront confondus avec ceux que (s'ils existent) nous avons construits en cherchant les diverses propriétés d'optimalité *exacte* (comparer avec [50]).

Le test π de choix entre H_1 et H_2 sera ici une fonction $\pi = \pi(X, Y)$ de deux échantillons X et Y qui, comme dans le chapitre 3, désignera la probabilité d'accepter H_2 pour la réunion des échantillons X et Y ou *échantillon global* (X, Y). Les définitions du niveau asymptotique et de l'optimalité asymptotique du test π sont les mêmes que dans le § 3.14.

DÉFINITION 1. On dira qu'un test π *est de niveau asymptotique* $1 - \epsilon$ (est de classe $\bar{K}_\epsilon$) si

$$\lim_{n\to\infty}\sup \sup_{(\theta_1,\theta_2)\in\bar{\Theta}_1} \mathbf{E}_{\theta_1\theta_2}\pi(X, Y) \leqslant \epsilon,$$

où $\mathbf{E}_{\theta_1\theta_2}$ est l'espérance mathématique par rapport à la distribution $\mathbf{P}_{\theta_1} \times \times \mathbf{P}_{\theta_2}$, et $\bar{\Theta}_1$ l'ensemble des valeurs (θ_1, θ_2) pour lesquelles est réalisée l'hypothèse H_1 (par exemple, l'ensemble de tous les points (θ_1, θ_2) situés sur la « bissectrice » $\theta_1 = \theta_2$ dans le problème d'homogénéité ordinaire).

DÉFINITION 2. On dit qu'un test $\pi_1 \in \bar{K}_\epsilon$ est *asymptotiquement minimax dans* $\bar{K}_\epsilon$ entre H_1 et H_2 si pour tout autre test $\pi \in \bar{K}_\epsilon$ on a

$$\lim_{n\to\infty}\inf \left(\inf_{(\theta_1,\theta_2)\in\bar{\Theta}_2} \mathbf{E}_{\theta_1\theta_2}\pi_1(X, Y) - \inf_{(\theta_1,\theta_2)\in\bar{\Theta}_2} \mathbf{E}_{\theta_1\theta_2}\pi(X, Y)\right) \geqslant 0,$$

où $\bar{\Theta}_2$ est l'ensemble des valeurs (θ_1, θ_2) correspondant aux alternatives H_2.

---

*) La recherche de solutions optimales a fait l'objet de nombreux travaux. Y. Linnik et son école ont apporté une importante contribution à l'étude du problème de Berens-Fisher qui est assez compliqué. Ces recherches impliquent de nouvelles notions et l'utilisation d'un outil mathématique assez complexe. D'où l'impossibilité de citer et de démontrer (dans le cadre de cet ouvrage) les résultats acquis. La situation se présente sous de meilleurs auspices dans les problèmes d'homogénéité ordinaire et d'homogénéité en présence d'un paramètre fantôme pour des populations normales (dans de nombreux problèmes on arrive à trouver des tests invariants uniformément les plus puissants sans biais, mais les constructions exigées sont assez compliquées : pour plus de détails voir [50]).

2. Test asymptotiquement minimax entre hypothèses voisines d'homogénéité ordinaire. Introduisons un nouveau paramètre $\bar{\theta} = (\theta_1, \theta_2)$ caractérisant la réunion des échantillons X et Y, dite encore échantillon global (X, Y). La fonction de vraisemblance de l'échantillon global (X, Y) est égal à $f_{\bar{\theta}}(X, Y) = f_{\theta_1}(X) f_{\theta_2}(Y)$.

Supposons par souci de simplicité que les échantillons sont de même taille : $n_1 = n_2 = n$. L'échantillon (X, Y) peut alors être représenté comme un échantillon de taille n formé par les couples d'observations (x_1, y_1), ..., (x_n, y_n) de distribution $\mathbf{P}_{\bar{\theta}} = \mathbf{P}_{\theta_1} \times \mathbf{P}_{\theta_2}$ et de densité $f_{\theta_1}(x) f_{\theta_2}(y)$. Nous sommes conduits au problème, envisagé dans le § 3.15, de test, au vu de l'échantillon (X, Y), de l'hypothèse H_1 que le paramètre $\bar{\theta}$ est situé sur la « courbe » $\theta_1 = \theta_2$. Si l'on adopte les notations du § 3.15, l'hypothèse H_1 s'écrit $H_1 = \{\bar{\theta} = g(\alpha)\}$, où $\alpha \equiv \theta_1$, $g(\alpha) = (\alpha, \alpha)$. Il est évident que la matrice $G = \left\| \frac{\partial g_i}{\partial \alpha_j} \right\|$, $i = 1, ..., 2k$, $j = 1, ..., k$, est de la forme $\begin{pmatrix} E \\ E \end{pmatrix}$, où E est la matrice unité d'ordre k, de sorte que rang $G = k$.

On admettra que la paramètre $\bar{\theta}$ est localisé, c'est-à-dire que les valeurs θ_1 et θ_2 sont proches et par suite les valeurs éventuelles de $\bar{\theta}$ sont situées dans un voisinage du point $\bar{\theta}_0 = (\theta_0, \theta_0)$ pour un θ_0 fixe. Si l'on suit le § 3.15, il nous sera plus commode d'introduire un nouveau paramètre $\tau = (\tau', \tau'') = (\gamma'/\sqrt{n}, \gamma''/\sqrt{n}) = \gamma/\sqrt{n}$, où $\tau' = \theta_1 - \theta_0$, $\tau'' = \theta_2 - \theta_1$, de sorte que l'application $\bar{\theta} = \bar{\theta}(\tau)$ est bijective : $\theta_1 = \tau' + \theta_0$, $\theta_2 = \tau'' + \tau' + \theta_0$. En termes de paramètres τ et γ, l'hypothèse H_1 d'homogénéité devient $H_1 = \{\tau'' = 0\} = \{\gamma'' = 0\}$. Pour hypothèse alternative nous considérons l'hypothèse « séparée »

$$H_2^b = \{\gamma'' I \gamma''^T \geqslant b^2\}, \quad b > 0, \tag{2}$$

où $I = I(\theta_0)$ est la matrice de Fisher pour la famille $\{\mathbf{P}_\theta\}$ au point θ_0.

THÉORÈME 1. *Supposons que la famille $\{\mathbf{P}_\theta\}$ vérifie les conditions (RR) au voisinage du point θ_0 (cf. § 2.28). Le test du rapport de vraisemblance*

$$R_1(X, Y) = \frac{\sup\limits_{\theta_1, \theta_2} f_{\theta_1}(X) f_{\theta_2}(Y)}{\sup\limits_{\theta} f_\theta(X) f_\theta(Y)} > e^{h_\epsilon/2} \tag{3}$$

est alors un test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$ de $H_1 = \{\theta_1 = \theta_2\}$ contre $H_2^b = \{(\theta_1 - \theta_2) I(\theta_1 - \theta_2)^T \geqslant b^2/n\}$ pour tout $b > 0$, où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à k degrés

de liberté (la statistique $2 \ln R_1(X, Y)$ *admet la même distribution limite pour l'hypothèse* H_1*).*

Supposons que $\hat{\theta}^*_X$, $\hat{\theta}^*_Y$, $\hat{\theta}^*$ *sont des estimateurs du maximum de vraisemblance du paramètre* $\theta = \theta_1 = \theta_2$ *au vu respectivement des échantillons* X, Y *et* (X, Y). *Le test*

$$(\hat{\theta}^*_X - \hat{\theta}^*)\, I(\hat{\theta}^*)(\hat{\theta}^*_X - \hat{\theta}^*)^T + (\hat{\theta}^*_Y - \hat{\theta}^*)\, I(\hat{\theta}^*)(\hat{\theta}^*_Y - \hat{\theta}^*)^T > h_\varepsilon/n \qquad (4)$$

sera alors asymptotiquement équivalent au test (3).

DÉMONSTRATION. Cette proposition est conséquence immédiate du théorème 3.15.4. Il nous faut seulement voir ce que sont la matrice de Fisher $\bar{I}(\bar{\theta}_0) = \bar{I}(\theta_0, \theta_0)$ pour le paramètre global $\bar{\theta} = (\theta_1, \theta_2)$ et la matrice M_2 pour la famille paramétrique $\{\mathbf{P}_{(\theta_0, \theta_0 + \beta)}\}$ au point $\beta = 0$. On a

$$\ln f_{\theta_1}(x) f_{\theta_2}(y) = l(x, \theta_1) + l(y, \theta_2).$$

Désignons par t_i, $i = 1, \ldots, 2k$, les coordonnées du vecteur $\bar{\theta}$. Si l'on désigne par $\mathbf{E}_{\bar{\theta}}$ l'espérance mathématique par rapport à la distribution $\mathbf{P}_{\bar{\theta}}$, les éléments $\bar{I}_{ij}(\bar{\theta})$ de la matrice $\bar{I}(\bar{\theta})$ seront alors égaux à

$$\bar{I}_{ij}(\bar{\theta}) = \mathbf{E}_{\bar{\theta}} \left(\frac{\partial l(\mathrm{x}_1, \theta_1)}{\partial t_i} + \frac{\partial l(\mathrm{y}_1, \theta_2)}{\partial t_i} \right) \left(\frac{\partial l(\mathrm{x}_1, \theta_1)}{\partial t_j} + \frac{\partial l(\mathrm{y}_1, \theta_2)}{\partial t_j} \right).$$

On en déduit en vertu de l'indépendance de x_1 et y_1 que

$$\bar{I}(\bar{\theta}) = \begin{pmatrix} I(\theta_1) & 0 \\ 0 & I(\theta_2) \end{pmatrix}.$$

Donc, le test (4) n'est autre que le test (3.15.12) du théorème 3.15.4.

Des calculs identiques nous montrent que $M_2 = I(\theta_0)$, puisque pour $\beta = (\beta_1, \ldots, \beta_k) = 0$

$$\frac{\partial l(\mathrm{x}_1, \theta_0)}{\partial \beta_i} + \frac{\partial l(\mathrm{y}_1, \theta_0 + \beta)}{\partial \beta_i} = \frac{\partial l(\mathrm{y}_1, \theta_0)}{\partial t_i}, \qquad i = 1, \ldots, k. \blacktriangleleft$$

REMARQUE 1. Nous avons prouvé le théorème 1 sous l'hypothèse que $n_1 = n_2$. Mais cette restriction n'est pas essentielle du tout. Considérons par exemple le cas où $n_1 \to \infty$ et $n_2 \to \infty$ de telle sorte que le rapport n_1/n_2 soit égal au rationnel r_1/r_2 (r_1 et r_2 sont des entiers quelconques fixes, $n_i = nr_i$, $n \to \infty$). Introduisons encore le paramètre $\bar{\theta} = (\theta_1, \theta_2)$ et traitons l'échantillon global (X, Y) comme un échantillon de taille n d'éléments $(\mathrm{x}_1, \ldots, \mathrm{x}_{r_1}; \mathrm{y}_1, \ldots, \mathrm{y}_{r_2})$, $(\mathrm{x}_{r_1+1}, \ldots, \mathrm{x}_{2r_1}; \mathrm{y}_{r_2+1}, \ldots, \mathrm{y}_{2r_2})$, ... dont la loi

$$\mathbf{P}_{\bar{\theta}} = \underbrace{\mathbf{P}_{\theta_1} \times \ldots \times \mathbf{P}_{\theta_1}}_{r_1 \text{ fois}} \times \underbrace{\mathbf{P}_{\theta_2} \times \ldots \times \mathbf{P}_{\theta_2}}_{r_2 \text{ fois}},$$

dépend du paramètre $\bar{\theta}$. La fonction de vraisemblance sera encore de la forme

$$f_\theta(X, Y) = f_{\theta_1}(X) f_{\theta_2}(Y).$$

Si l'on introduit comme précédemment le paramètre $\tau = (\tau', \tau'') = (\theta_1 - \theta_0, \theta_2 - \theta_1)$ et que l'on pose $\tau = \gamma/\sqrt{n} = (\gamma'/\sqrt{n}, \gamma''/\sqrt{n})$, le problème posé consiste à tester l'hypothèse $H_1 = \{\gamma'' = 0\}$ contre l'hypothèse $H_2^b = \{\gamma'' M_2 \gamma''^T \geqslant b^2\}$, où M_2 est la matrice de Fisher pour $\mathbf{P}_{(\theta_0, \theta_0 + \beta)}$ au point $\beta = 0$. Il est aisé de voir que dans notre cas $M_2 = r_2 I(\theta_0)$, de sorte que l'ensemble des alternatives conserve sa forme (2) :

$$H_2^b = \{\gamma'' I \gamma''^T > b^2/r^2\}.$$

La matrice de Fisher $\bar{I}(\theta)$ devient

$$\begin{pmatrix} r_1 I(\theta_1) & 0 \\ 0 & r_2 I(\theta_2) \end{pmatrix}.$$

Reste à appliquer le théorème 3.15.4. Nous obtenons alors la proposition du théorème 1 dans laquelle il faut remplacer le test (4) par

$$n(\delta_X^* - \theta^*) I(\theta^*)(\delta_X^* - \theta^*)^T + n_2(\delta_Y^* - \theta^*) I(\theta^*)(\delta_Y^* - \theta^*) > h_\epsilon. \quad (5)$$

Le théorème 3.15.4 nous permet de déterminer aussi la puissance asymptotique garantie des tests (3), (4) et (5).

Ce théorème reste valable dans le cas général où $n_1 \to \infty$, $n_2 \to \infty$, $n_1/n_2 \to c$, où c est un nombre arbitraire de $]0, 1[$. Mais la démonstration de ce fait implique des considérations supplémentaires.

REMARQUE 2. Le théorème 1 reste en vigueur si l'on remplace l'hypothèse $H_1 = \{\theta_1 = \theta_2\}$ par

$$H_1^a = \{(\theta_1 - \theta_2) I(\theta_1 - \theta_2)^T \leqslant a^2/n\}, \qquad 0 < a < b.$$

REMARQUE 3. La forme des tests asymptotiquement minimax du théorème 1 ne dépend pas de θ_0. La valeur θ_0 ne figure dans la définition de l'hypothèse H_2^b que par l'intermédiaire de $I = I(\theta_0)$ (cf. (2)). On aurait pu éviter l'apparition de θ_0 en remplaçant I par $I((\theta_1 + \theta_2)/2)$ dans (2). Ceci nous aurait fourni une hypothèse $\tilde{H}_2^b$ « asymptotiquement équivalente » à H_2^b pour laquelle le théorème 3 reste entièrement en vigueur. L'apparition de θ_0 dans (2) est la conséquence de l'utilisation d'une méthode plus simple de réduction du problème envisagé aux résultats du § 3.15.

EXEMPLE 1. Supposons que X et Y sont des échantillons de taille n_1 et n_2 distribués respectivement suivant les lois polynomiales $\mathbf{B}_{\theta_1}$ et $\mathbf{B}_{\theta_2}$, $\theta_i \in R^k$, $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})$, $i = 1, 2$. Les vecteurs $\nu = (\nu_1, \ldots, \nu_k)$ et $\mu = (\mu_1, \ldots, \mu_k)$

des fréquences d'apparitions des événements $A_1, \ldots, A_k$ (cf. § 2.2) forment les statistiques exhaustives

$$f_{\theta_1}(X) = \prod_{i=1}^{k} \theta_{1i}^{\nu_i}, \quad f_{\theta_2}(Y) = \prod_{i=1}^{k} \theta_{2i}^{\mu_i}.$$

Les estimateurs du maximum de vraisemblance sont de la forme $\hat{\theta}_X^* = \nu/n_1$, $\hat{\theta}_Y^* = \mu/n_2$, $\hat{\theta}^* = (\nu + \mu)/(n_1 + n_2)$. La matrice $I(\theta)$ a été définie dans (3.16.5), de sorte que (cf. (3.16.9))

$$tI(\theta_0)t^T = \sum_{i=1}^{k} \frac{t_i^2}{\theta_{0i}}.$$

Ainsi, en vertu du théorème 1 et de la remarque 1, le test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$ de $H_1 = \{\theta_1 = \theta_2\}$ contre

$$H_2^b = \left\{ \sum_{i=1}^{k} (\theta_{1i} - \theta_{2i})^2/\theta_{0i} \geqslant b^2/n_2 \right\}$$

est de la forme

$$\ln R_1(X, Y) =$$

$$= \sum_{i=1}^{k} \nu_i \ln \frac{\nu_i}{n_1} + \sum_{i=1}^{k} \mu_i \ln \frac{\mu_i}{n_2} - \sum (\nu_i + \mu_i) \ln \frac{\nu_i + \mu_i}{n_1 + n_2} > \frac{h_\epsilon}{2},$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à $k - 1$ degrés de liberté. D'après (4) et (5), le test

$$n_1 \sum_{i=1}^{k} \left(\frac{\nu_i}{n_1} - \frac{\nu_i + \mu_i}{n_1 + n_2}\right)^2 \frac{n_1 + n_2}{\nu_i + \mu_i} +$$

$$+ n_2 \sum_{i=1}^{k} \left(\frac{\mu_i}{n_2} - \frac{\nu_i + \mu_i}{n_1 + n_2}\right)^2 \frac{n_1 + n_2}{\nu_i + \mu_i} =$$

$$= \sum_{i=1}^{k} \left(\frac{\nu_i}{n_1} - \frac{\mu_i}{n_2}\right)^2 \frac{n_1 n_2}{\nu_i + \mu_i} > h_\epsilon \quad (6)$$

lui sera asymptotiquement équivalent.

EXEMPLE 1A. Dans l'exemple 2.26.3 nous avons décrit le mécanisme de transmission des groupes sanguins 0, A, B, AB. Ce mécanisme est commandé par les gènes A, B et 0. Désignons respectivement par p, q et $r = 1 - p - q$ les probabilités d'apparition de ces gènes dans une population donnée. Les probabilités $p_i(\alpha)$, $\alpha = (p, q)$, qu'un individu soit du groupe i s'expriment en fonction de α à l'aide des formules du tableau 1 du § 2.26.

On dispose de deux échantillons X et Y de fréquences respectives ν_i et μ_i d'apparition du groupe $i = 1, \ldots, 4$, obtenus par un sondage effectué sur $n_1 = 353$ personnes de la communauté I et $n_2 = 364$ personnes de la communauté II. Les résultats sont consignés dans le tableau 1.

Tableau 1

communauté \ groupe sanguin	0	A	B	AB	Total
I	121	120	79	33	353
II	118	95	121	30	364
Total	239	215	200	63	717

Il faut tester l'hypothèse que les communautés sondées appartiennent à une même population, c'est-à-dire l'hypothèse que les probabilités p et q sont égales pour ces groupes, ou ce qui est équivalent que les $p_i(\alpha)$ sont égales. On reconnaît de toute évidence le problème d'homogénéité étudié dans l'exemple 1.

Si l'on teste la coïncidence des probabilités des quatre groupes sanguins, la distribution limite de la statistique (cf. (6))

$$\chi_1^2 = \sum_{i=1}^{4} \left(\frac{\nu_i}{n_1} - \frac{\mu_i}{n_2}\right)^2 \frac{n_1 n_2}{\nu_i + \mu_i}$$

sera la distribution du χ^2 à trois degrés de liberté. Dans notre cas, $\chi_1^2 = 11,74$. Le niveau réel (cf. § 3.4) de l'écart obtenu est supérieur à 0,99. Ceci exprime que l'hypothèse d'homogénéité doit être infirmée par le test $\chi_1^2 > h_{0,01}$ de niveau 0,99.

A noter que le test utilisé ne correspond pas entièrement à la nature de l'événement étudié, puisque nous devons éprouver la coïncidence des probabilités p et q et pas des probabilités p_i d'apparition des groupes sanguins. Si l'on suit exactement le théorème 1, on doit calculer, à l'aide des méthodes du § 2.26, les estimateurs du maximum de vraisemblance $\hat{\alpha}_X^*$, $\hat{\alpha}_Y^*$ et $\hat{\alpha}^*$

du paramètre $\alpha = (p, q)$ au vu respectivement des échantillons X, Y et (X, Y) et utiliser la statistique

$$\chi_2^2 = 2[L(\hat{\alpha}_X^*, X) + L(\hat{\alpha}_Y^*, Y) - L(\hat{\alpha}^*, (X, Y))] =$$

$$= 2\left[\sum_{i=1}^{4} \nu_i \ln p_i(\hat{\alpha}_X^*) + \sum_{i=1}^{4} \mu_i \ln p_i(\hat{\alpha}_Y^*) - \sum_{i=1}^{4} (\nu_i + \mu_i) \ln p_i(\hat{\alpha}^*)\right]$$

qui, pour les grands n, est distribuée suivant une loi proche de celle du χ^2 à deux degrés de liberté. Si l'on effectue les calculs nécessaires (cf. exemple 2.26.3), on obtient $\chi_2^2 \approx 11{,}04$, ce qui, pour deux degrés de liberté, donne un écart plus significatif que 11,74 pour trois.

Le test de l'hypothèse que X et Y sont distribués suivant des lois appartenant aux sous-familles paramétriques $\mathbf{B}_{p(\alpha)}$, où $p(\alpha) = (p_1(\alpha), \ldots$ $\ldots, p_4(\alpha))$, est étudié dans l'exemple 3.17.1. Les deux échantillons s'accordent bien avec cette hypothèse.

EXEMPLE 2. Soient $X \in \Phi_{\alpha_1, \sigma_1^2}$, $Y \in \Phi_{\alpha_2, \sigma_2^2}$, où les points $\theta_i = (\alpha_i, \sigma_i^2)$ sont situés au voisinage du point $\theta_0 = (\alpha_0, \sigma_0^2)$. On a

$$I(\theta_0) = \begin{pmatrix} \sigma_0^{-2} & 0 \\ 0 & \frac{1}{2}\sigma_0^{-4} \end{pmatrix}$$

(cf. § 2.16) et l'on envisagera le problème de choix entre l'hypothèse $H_1 =$ $= \{\theta_1 = \theta_2\}$ et

$$H_2^b = \left\{\frac{(\alpha_1 - \alpha_2)^2}{\sigma_0^2} + \frac{(\sigma_1^2 - \sigma_2^2)^2}{2\sigma_0^4} \geq \frac{b^2}{n_2}\right\}, \quad n = n_1 + n_2.$$

On a $\theta_X^* = (\bar{x}, S_X^2), S_X^2 = \frac{1}{n_1}\sum_{i=1}^{n_1} (x_i - \bar{x})^2$, $f_{\theta_X^*}(X) = (2\pi e S_X^2)^{-n_1/2}$.

L'échantillon Y est justiciable des mêmes formules. Par ailleurs,

$$\theta^* = (\bar{z}, S_{X,Y}^2),\ \bar{z} = \frac{\left(\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i\right)}{n_1 + n_2} = a\bar{x} + (1 - a)\bar{y},$$

$$S_{X,Y}^2 = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} (x_i - \bar{z})^2 + \sum_{i=1}^{n_2} (y_i - \bar{z})^2\right] = \tag{7}$$

$$= aS_X^2 + (1 - a) S_Y^2 + (1 - a) a(\bar{x} - \bar{y})^2,$$

où $a = n_1/(n_1 + n_2)$, $f_{\hat\theta}.(X) f_{\hat\theta}.(Y) = (2\pi e S^2_{X,Y})^{-\frac{1}{2}(n_1+n_2)}$. Donc le test

$$\frac{S^2_{X,Y}}{S_X^{2a}\, S_Y^{2(1-a)}} > e^{h_\epsilon/(n_1+n_2)},$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à deux degrés de liberté, est un test asymptotiquement minimax entre H_1 et H_2^b. Nous proposons au lecteur de trouver à titre d'exercice un test asymptotiquement équivalent de la forme (5).

3. Tests asymptotiquement minimax pour le problème d'homogénéité en présence d'un paramètre fantôme. Dans ce numéro et les suivants, on admettra pour simplifier que les échantillons X et Y sont de même taille : $n_1 = n_2$. Cette restriction n'est pas essentielle. Si $n_1/n_2 = r_1/r_2$ (r_1 et r_2 étant des entiers), le lecteur pourra procéder comme dans la remarque 1 suivant le théorème 1.

Soient donc donnés deux échantillons $X \in \mathbf{P}_{\theta_1}$ et $Y \in \mathbf{P}_{\theta_2}$, $\theta_i = (u_i, v_i)$, $i = 1, 2$, de taille $n_1 = n_2 = n$. On teste l'hypothèse $\{u_1 = u_2\}$ contre $\{u_1 \neq \neq u_2\}$ sous la condition que $v_1 = v_2 = v$ et v est inconnu. Désignons la dimension de u_i par l, $l < k$.

Introduisons un nouveau paramètre $\bar\theta = (u_1, u_2, v)$. Représentons l'échantillon global (X, Y) comme un échantillon de taille n d'éléments $(\mathrm{x}_1, \mathrm{y}_1), \ldots, (\mathrm{x}_n, \mathrm{y}_n)$ dont la densité de probabilité est égale à $f_{\bar\theta}(x, y) = = f_{(u_1,v)}(x) f_{(u_2,v)}(y)$. Pour cette famille paramétrique, le problème envisagé est équivalent au problème de test de l'hypothèse H_1 que la valeur $\bar\theta$ est située sur la « courbe » $\bar\theta = g(\theta_1) = (u_1, u_1, v)$. La matrice $G = \left\| \frac{\partial g_i}{\partial \theta_{1j}} \right\|$, $i = 1, \ldots, k + l$, $j = 1, \ldots, k$, est de la forme $\begin{pmatrix} E_l & 0 \\ & E_k \end{pmatrix}$, où E_l est la matrice unité d'ordre l et E_k la matrice unité d'ordre k, de sorte que rang $G = k$.

Comme dans le numéro précédent nous conviendrons que le paramètre θ est localisé au voisinage du point $\theta_0 = (u_0, v_0)$. Introduisons le paramètre $\tau = \tau(\bar\theta) = (\tau', \tau'', \tau''') = (u_1 - u_0, u_2 - u_1, v - v_0)$. La contre-image $\bar\theta = \bar\theta(\tau)$ existe toujours et possède les coordonnées $u_1 = \tau' + u_0$, $u_2 = = \tau'' + \tau' + u_0$, $v = \tau''' + v_0$. Posons $\tau = \gamma/\sqrt{n}$, $\gamma = (\gamma', \gamma'', \gamma''')$.

Pour le nouveau paramètre τ (ou γ) l'hypothèse d'homogénéité s'écrit $H_1 = \{\gamma'' = 0\}$. Pour hypothèse alternative, considérons l'hypothèse « séparée » $H_2^b = \{\gamma'' I_1(\theta_0) \gamma''^{t} \geqslant b^2\}$, où $I_1(\theta)$ est la matrice formée par les l premières lignes et colonnes de la matrice initiale d'information de Fisher $I(\theta)$.

THÉORÈME 2. *Supposons que la famille $\{\mathbf{P}_\theta\}$ remplit les conditions (RR) au voisinage du point θ_0. Le test du rapport de vraisemblance*

•

$$R_1(X, Y) = \frac{\sup\limits_{(u_1, u_2, v)} f_{(u_1, v)}(X) f_{(u_2, v)}(Y)}{\sup\limits_{\theta} f_\theta(X) f_\theta(Y)} > e^{h_\epsilon/2} \tag{8}$$

est alors un test asymptotiquement minimax de niveau asymptotique $1 - \epsilon$ *de* $H_1 = \{u_1 = u_2\}$ *contre*

$$H_2^b = \{(u_1 - u_2) I_1(\theta_0)(u_1 - u_2)^T \geqslant b^2/n\} \tag{9}$$

pour tout $b > 0$ *et* $v_1 = v_2 = v$. *Ici* h_ϵ *est le quantile d'ordre* $1 - \epsilon$ *de la distribution du* χ^2 *à* l *degrés de liberté. (Cette distribution sera distribution limite de la statistique* $2 \ln R_1(X, Y)$ *pour l'hypothèse* H_1.)

Désignons par $\bar{\theta}^*$ *et* $\theta^* = (u^*, v^*)$ *les valeurs des paramètres* $\bar{\theta}$ *et* θ *pour lesquelles est réalisé le maximum respectivement du numérateur et du dénominateur de* (8). *Mettons la matrice* $I(\theta)$ *sous la forme*

$$I(\theta) = \begin{pmatrix} I_1(\theta) & I_{21}(\theta) \\ I_{12}(\theta) & I_{22}(\theta) \end{pmatrix}.$$

Le test

$$(\bar{\theta}^* - (u^*, u^*, v^*)) \bar{I}(\bar{\theta}^*)(\bar{\theta}^* - (u^*, u^*, v^*))^T > h_\epsilon/n, \tag{10}$$

où

$$\bar{I}(\bar{\theta}) = \begin{pmatrix} I_1(\theta_1) & 0 & I_{21}(\theta_1) \\ 0 & I_1(\theta_2) & I_{21}(\theta_2) \\ I_{12}(\theta_1) & I_{12}(\theta_2) & I_{22}(\theta_1) + I_{22}(\theta_2) \end{pmatrix}, \tag{11}$$

sera alors asymptotiquement équivalent à (8).

DÉMONSTRATION. Ce théorème est conséquence directe du théorème 3.15.4. Il nous faut seulement déterminer la structure de la matrice $\bar{I}(\bar{\theta})$ pour l'échantillon (X, Y) et le paramètre « global » $\bar{\theta}$, et celle de la matrice M_2. On a

$$l \equiv \ln f_\theta(x, y) = l(x, (u_1, v)) + l(y, (u_2, v)).$$

Désignons par t_i, $i = 1, \ldots, k + l$, les coordonnées du vecteur $\bar{\theta}$. Alors

$$\frac{\partial l}{\partial t_i} = \begin{cases} \dfrac{\partial l(x, (u_1, v))}{\partial t_i}, & 0 < i \leqslant l, \\[2ex] \dfrac{\partial l(y, (u_2, v))}{\partial t_i}, & l < i \leqslant 2l, \\[2ex] \dfrac{\partial l(x, (u_1, v))}{\partial t_i} + \dfrac{\partial l(y, (u_2, v))}{\partial t_i}, & 2l < i \leqslant k + l; \end{cases}$$

d'où l'on déduit (11) sans peine.

La matrice M_2 pour la famille paramétrique $\mathbf{P}_{\theta(0,\beta,0)} = \mathbf{P}_{(u_0,u_0+\beta,v_0)}$ se calcule de façon analogue au point $\beta = 0$, est égale à $I_1(\theta_0)$ et correspond à la sous-matrice moyenne de la matrice $\overline{I}(\overline{\theta}_0)$. ◄

Dans les exemples suivants les tailles n_1 et n_2 des échantillons sont arbitraires.

EXEMPLE 3. Soient $X \in \Phi_{\alpha_1,\sigma^2}$ et $Y \in \Phi_{\alpha_2,\sigma^2}$. On demande de tester l'hypothèse $H_1 = \{\alpha_1 = \alpha_2\}$, σ^2 étant inconnue. Pour déterminer les tests asymptotiquement minimax à l'aide du théorème 2, il nous faut trouvèr la statistique $R_1(X, Y)$ de (8), où, dans cet exemple, $u_i = \alpha_i$, $v = \sigma^2$, $\overline{\theta} = (\alpha_1, \alpha_2, \sigma^2)$. On a $\ln f_{(\alpha_1,\sigma^2)}(X) f_{(\alpha_2,\sigma^2)}(Y) = -\frac{1}{2}(n_1 + n_2)\ln(2\pi\sigma^2) -$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i - \alpha_1)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_2}(y_i - \alpha_2)^2.$$ En annulant les dérivées de cette fonction par rapport à α_1, α_2 et σ^2, et en résolvant les équations obtenues, on trouve (dans les notations de l'exemple 2)

$$\begin{aligned} &\overline{\theta}^* = (\overline{x}, \overline{y}, aS_X^2 + (1-a)S_Y^2), \quad a = \frac{n_1}{n_1+n_2}, \\ &f_{\overline{\theta}^*}(X, Y) = [2\pi e(aS_X^2 + (1-a)S_Y^2)]^{-(n_1+n_2)/2}. \end{aligned} \tag{12}$$

En procédant de même avec la fonction $\ln f_\theta(X) f_\theta(Y) = \ln f_{(\alpha,\sigma^2)}(X) f_{(\alpha,\sigma^2)}(Y)$, on obtient (cf. exemple 2)

$$\begin{aligned} &\theta^* = (\overline{z}, S_{X,Y}^2), \\ &f_{\theta^*}(X) f_{\theta^*}(Y) = (2\pi e S_{X,Y}^2)^{-\frac{1}{2}(n_1+n_2)}. \end{aligned} \tag{13}$$

Un test asymptotiquement optimal sera donc de la forme

$$\frac{S_{X,Y}^2}{aS_X^2 + (1-a)S_Y^2} > e^{h_\epsilon/(n_1+n_2)},$$

ou (cf. (7))

$$\frac{\sqrt{a(1-a)}\,|\overline{x} - \overline{y}|}{\sqrt{aS_X^2 + (1-a)S_Y^2}} > \sqrt{\frac{h_\epsilon}{n_1+n_2}},$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à un degré de liberté, si bien que $\sqrt{h_\epsilon}$ peut être remplacée par la valeur $\lambda_{\epsilon/2}$ pour laquelle $\Phi_{0,1}(]-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}[) = 1 - \epsilon$. Il est immédiat de voir que le premier membre

de l'inégalité

$$\frac{\sqrt{a(1-a)(n_1+n_2)}\,|\bar{x}-\bar{y}|}{\sqrt{aS_X^2+(1-a)S_Y^2}} > \lambda_{\epsilon/2} \tag{14}$$

qui définit un test asymptotiquement minimax, sera une variable aléatoire asymptotiquement normale de paramètres (0, 1) après la substitution de $\bar{x}-\bar{y}$ à $|\bar{x}-\bar{y}|$.

Mais ce test peut être rendu exact (c'est-à-dire de niveau exact donné à l'avance). En effet, en vertu des résultats du § 2.32 pour l'hypothèse H_1

$$\sqrt{\frac{n_1 n_2}{n_1+n_2}}\,\frac{\bar{x}-\bar{y}}{\sigma} \in \Phi_{0,1},$$

$$\frac{(n_1+n_2)\,aS_X^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(x_i-\bar{x})^2 \in \mathbf{H}_{n_1-1},$$

$$\frac{(n_1+n_2)(1-a)\,S_Y^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n_2}(y_i-\bar{y})^2 \in \mathbf{H}_{n_2-1}.$$

Ces trois variables aléatoires étant indépendantes, le rapport

$$(\bar{x}-\bar{y})\sqrt{\frac{n_1 n_2}{n_1+n_2}}\left[\frac{n_1+n_2}{n_1+n_2-2}\,(aS_X^2+(1-a)\,S_Y^2)\right]^{-1/2} =$$
$$= \frac{(\bar{x}-\bar{y})\sqrt{a(1-a)(n_1+n_2-2)}}{\sqrt{aS_X^2+(1-a)\,S_Y^2}} \in \mathbf{T}_{n_1+n_2-2}$$

sera distribué suivant la loi de Student à n_1+n_2-2 degrés de liberté. Donc, le test (comparer avec (14))

$$\frac{(\bar{x}-\bar{y})\sqrt{a(1-a)(n_1+n_2-2)}}{\sqrt{aS_X^2+(1-a)\,S_Y^2}} > \tau_\epsilon,$$

où τ_ϵ est tel que $\mathbf{T}_{n_1+n_2-2}(]-\tau_\epsilon, \tau_\epsilon[) = 1-\epsilon$, aura un niveau exactement égal à $1-\epsilon$ et on peut l'utiliser pour n'importe quelles valeurs de n_1 et n_2 (et pas seulement pour les grandes). Ce test est appelé *test de Student*. Il possède aussi certaines propriétés d'optimalité exacte et pas seulement asymptotique (cf. [50]).

EXEMPLE 4. Soient $X \in \Phi_{\alpha,\sigma_1^2}$ et $Y \in \Phi_{\alpha,\sigma_2^2}$. On teste l'hypothèse $\{\sigma_1 = \sigma_2\}$ pour α inconnu. En procédant comme dans l'exemple précédent, on

aboutit à une statistique (8) dont le dénominateur est le même que dans l'exemple précédent et le numérateur égal à

$$\sup_{(\alpha,\sigma_1,\sigma_2)} f_{(\alpha,\sigma_1^2)}(X) f_{(\alpha,\sigma_2^2)}(Y). \tag{15}$$

Ecrivons les équations du point de maximum

$$\sigma_1^2 = \frac{1}{n_1} \sum_{i=1}^{n_1} (x_i - \alpha)^2 = S_X^2 + (\bar{x} - \alpha)^2,$$

$$\sigma_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (y_i - \alpha)^2 = S_Y^2 + (\bar{y} - \alpha)^2,$$

$$\frac{n_1}{\sigma_1^2}(\bar{x} - \alpha) + \frac{n_2}{\sigma_2^2}(\bar{y} - \alpha) = 0.$$

En posant

$$p = \frac{a}{\sigma_1^2} \cdot \frac{1}{a/\sigma_1^2 + (1 - \alpha)/\sigma_2^2} \in]0, 1[, \tag{16}$$

on en déduit que

$$\alpha = p\bar{x} + (1 - p)\,\bar{y},$$

$$\sigma_1^2 = S_X^2 + (1 - p)^2 \Delta^2, \quad \sigma_2^2 = S_Y^2 + p^2\Delta^2,$$

où, pour simplifier, on a posé $\Delta = \bar{x} - \bar{y}$; p peut être traité comme la solution de l'équation (16) ou

$$p = \frac{a(S_Y^2 + p^2\Delta^2)}{a(S_Y^2 + p^2\Delta^2) + (1 - a)(S_X^2 + (1 - p)^2\Delta^2)} \cdot$$

Vu que le maximum de (15) est égal à

$$(2\pi e)^{-(n_1+n_2)/2} (S_X^2 + (1 - p)^2\Delta^2)^{-n_1/2} (S_Y^2 + p^2\Delta^2)^{-n_2/2}, \tag{17}$$

en comparant cette expression à (13) et (7), on obtient le test asymptotiquement minimax

$$\frac{aS_X^2 + (1 - a)S_Y^2 + a(1 - a)\Delta^2}{(S_X^2 + (1 - p)^2\Delta^2)^a (S_Y^2 + p^2\Delta^2)^{1-a}} > e^{h_\varepsilon/(n_1+n_2)}, \tag{18}$$

ou

$$\frac{aS_X^2 + (1 - a)S_Y^2}{S_X^{2a} S_Y^{2(1-a)}} > e^{h_\varepsilon/(n_1+n_2)} A^{-1}, \tag{19}$$

où $A = \dfrac{1 + \dfrac{a(1-a)\Delta^2}{aS_X^2 + (1-a)S_Y^2}}{(1 + (1-p)^2\Delta^2/S_X^2)^a(1 + p^2\Delta^2/S_Y^2)^{1-a}}$, et h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution du χ^2 à un degré de liberté. Ici $\Delta^2 = (\sigma_1^2/n_1 + \sigma_2^2/n_2)\xi^2$, $\xi \in \Phi_{0,1}$, $S_X^2/\sigma_1^2 \xrightarrow[P]{} 1$, $S_Y^2/\sigma_2^2 \xrightarrow[P]{} 1$, $\sigma_1^2/\sigma_2^2 \to 1$, $p \to a$

(on peut admettre pour simplifier que $a = \dfrac{n_1}{n_1 + n_2}$ est fixé), $\ln A \xrightarrow[P]{} 0$ pour chacune des hypothèses voisines envisagées. Donc, le second membre de (19) est de la forme

$$1 + \frac{h_\epsilon + \delta_n}{n_1 + n_2}, \quad \delta_n \xrightarrow[P]{} 0.$$

Le premier membre de (19) est le rapport de la moyenne arithmétique à la moyenne géométrique des quantités S_X^2 et S_Y^2. Si l'on désigne S_X^2/S_Y^2 par Z^2, l'inégalité contraire de (19) peut être mise sous la forme

$$\frac{aZ^2 + (1-a)}{Z^{2a}} - 1 \leqslant \frac{h_\epsilon + \delta_n}{n_1 + n_2}. \tag{20}$$

Au premier membre figure une fonction de Z convexe vers le bas (on peut, pour fixer les idées, admettre que $a \leqslant 1/2$) présentant un zéro multiple en $Z = 1$. Vu que le second membre de cette inégalité est petit, il nous sera commode de chercher la solution sous la forme $Z^2 = 1 + \zeta$ pour ζ petit. En limitant le développement en série suivant les puissances de ζ à l'ordre deux, on obtient pour les bornes ζ_1 et ζ_2 de l'intervalle sur lequel est vérifiée (20) les valeurs

$$\zeta_1 = \sqrt{\frac{2(h_\epsilon + \delta_n')}{a(1-a)(n_1 + n_2)}}, \quad \zeta_2 = \sqrt{\frac{2(h_\epsilon + \delta_n'')}{a(1-a)(n_1 + n_2)}},$$
$$\delta_n' \xrightarrow[P]{} 0, \quad \delta_n'' \xrightarrow[P]{} 0.$$

Ceci exprime, si l'on revient aux variables de départ, que le domaine

$$\sqrt{\frac{1}{2}a(1-a)(n_1 + n_2)}\,|S_X^2/S_Y^2 - 1| > \sqrt{h_\epsilon} = \lambda_{\epsilon/2} \tag{21}$$

(λ_ϵ est défini dans l'exemple 3) définit un test asymptotiquement équivalent à (18) et par suite, asymptotiquement minimax.

Comme dans l'exemple 3 nous pouvons rendre le test obtenu exact, puisque l'on connaît la distribution exacte de la statistique S_X^2/S_Y^2. En effet

$$n_1 S_X^2/\sigma_1^2 \in \mathbf{H}_{n_1 - 1}, \quad n_2 S_Y^2/\sigma_2^2 \in \mathbf{H}_{n_2 - 1}$$

et pour l'hypothèse $H_1 = \{\sigma_1 = \sigma_2\}$

$$\frac{n_1 S_X^2}{n_2 S_Y^2} \in \mathbf{F}_{n_1 - 1, n_2 - 1},$$

où $\mathbf{F}_{n_1 - 1, n_2 - 1}$ est la distribution de Fisher introduite dans le § 2.2 et tabulée dans de nombreux ouvrages de statistique. Ceci exprime que l'on peut calculer le niveau de signification exact du test (21) et l'appliquer pour tous n_1 et n_2 (pour les propriétés d'optimalité exacte de ce test cf. [50]). Si n_1 et n_2 sont grands, le premier membre de (21) (considéré sans le signe de la valeur absolue) est une variable aléatoire asymptotiquement normale de paramètres (0, 1).

4. Test asymptotiquement minimax pour le problème d'homogénéité partielle. Soient $X \in \mathbf{P}_{\theta_1}$, $Y \in \mathbf{P}_{\theta_2}$, $\theta_i = (u_i, v_i)$, $i = 1, 2$. On teste l'hypothèse $\{u_1 = u_2\}$ contre $\{u_1 \neq u_2\}$ pour des v_1 et v_2 quelconques. Comme précédemment dim $u_i = l$, $l < k$.

Introduisons le nouveau paramètre $\bar{\theta} = (\theta_1, \theta_2) = (u_1, v_1, u_2, v_2)$ de dimension $2k$. Comme précédemment traitons l'échantillon (X, Y) (pour $n_1 = n_2 = n$) comme un échantillon d'éléments $(\mathrm{x}_1, \mathrm{y}_1), \ldots, (\mathrm{x}_n, \mathrm{y}_n)$ de densité

$$f_{\bar{\theta}}(x, y) = f_{(u_1, v_1)}(x) f_{(u_2, v_2)}(y).$$

Pour cette famille le problème d'homogénéité partielle est équivalent au problème de test de l'hypothèse que $\bar{\theta}$ est situé sur la « courbe » $\theta = g(\alpha) = (u_1, v_1, u_1, v_2)$, où $\alpha = (u_1, v_1, v_2)$ est un « sous-paramètre » de dimension $2k - l$. Nous proposons au lecteur de s'inspirer des raisonnements des numéros précédents pour écrire la matrice $G = \left\| \frac{\partial g_i}{\partial \alpha_j} \right\|$, $i = 1, \ldots, 2k$, $j = 1, \ldots, 2k - l$, dont le rang sera égal à $2k - l$.

Comme dans les numéros 2 et 3, nous admettrons que le problème est « localisé » au voisinage du point $\theta_0 = (u_0, v_0)$. Introduisons le paramètre $\tau = \tau(\bar{\theta}) = (\tau', \tau'', \tau''', \tau^{\mathrm{IV}}) = (u_1 - u_0, v_1 - v_0, u_2 - u_1, v_2 - v_0)$. La contre-image a pour coordonnées

$$u_1 = \tau' + u_0, \quad v_1 = \tau'' + v_0, \quad u_2 = \tau''' + \tau' + u_0,$$
$$v_2 = \tau^{\mathrm{IV}} + v_0.$$

Si l'on pose $\tau = \gamma/\sqrt{n}$, $\gamma = (\gamma', \gamma'', \gamma''', \gamma^{\mathrm{IV}})$, l'hypothèse H_1 sera de la forme $H_1 = \{\gamma''' = 0\}$. Pour hypothèse concurrente on considérera l'hypothèse « séparée » $H_2^b = \{\gamma''' I_1(\theta_0) \gamma'''^{T} \geqslant b^2\}$, où $I_1(\theta)$ admet la même signification que dans le théorème 2.

THÉORÈME 3. *Supposons que la famille* $\{\mathbf{P}_\theta\}$ *vérifie les conditions* (*RR*) *au voisinage du point* θ_0. *Le test du rapport de vraisemblance*

$$R_1(X, Y) = \frac{\sup\limits_{(\theta_1,\theta_2)} f_{\theta_1}(X) f_{\theta_2}(Y)}{\sup\limits_{(u,v_1,v_2)} f_{(u,v_1)}(X) f_{(u,v_2)}(Y)} > e^{h_\epsilon/2} \qquad (22)$$

est alors un test asymptotiquement minimax de niveau asymptotique $1-\epsilon$ *de* H_1 *contre* H_2^b *définie dans* (9) *pour* v_1 *et* v_2 *quelconques. La valeur* h_ϵ *est la même que dans le théorème* 2.

DÉMONSTRATION. Elle reprend les raisonnements des numéros précédents et est entièrement basée sur le théorème 3.15.4. La recherche de la matrice d'information de Fisher $\bar{I}(\bar{\theta})$ pour le paramètre $\bar{\theta}$ et de la matrice M_2 pour la famille de densité $f_{\theta(0,0,\beta,0)} = f_{(u_0,v_0,u_0+\beta,v_0)}$ au point $\beta = 0$ est laissée au soin du lecteur.

La matrice $\bar{I}((\theta_X^*, \theta_Y^*))$ et le vecteur $(\theta_X^*, \theta_Y^*) - (u^*, v_1^*, u^*, v_2^*)$, où (θ_X^*, θ_Y^*) et (u^*, v_1^*, v_2^*) sont les vecteurs réalisant le maximum du numérateur et du dénominateur de (22), nous permettent comme nous l'avons fait précédemment à l'aide du théorème 3.15.4 (cf. 3.15.12) de construire un test asymptotiquement équivalent utilisant la forme quadratique des estimateurs introduits. ◄

EXEMPLE 5. *Comparaison des variances des lois normales*. Soient $X \in \Phi_{\alpha_1,\sigma_1^2}$, $Y \in \Phi_{(\alpha_2,\sigma_2^2)}$, $H_1 = \{\sigma_1 = \sigma_2\}$. Les calculs sont bien plus aisés que dans l'exemple 4, car la valeur du numérateur de (22) (de même que le vecteur $(\theta_X^*, \theta_Y^*) = (\bar{x}, S_X^2, \bar{y}, S_Y^2)$) est connue et le dénominateur a été calculé dans l'exemple 3 (cf. (12)). L'inégalité (22) sera ici de la forme

$$\frac{aS_X^2 + (1-a)S_Y^2}{S_X^{2a} S_Y^{2(1-a)}} > e^{h_\epsilon/(n_1+n_2)}.$$

En comparant ceci à (19) et aux considérations postérieures, on est conduit aux mêmes tests et conclusions que dans l'exemple 4.

EXEMPLE 6. *Problème de Berens-Fisher de comparaison des moyennes de deux lois normales*. Soient $X \in \Phi_{\alpha_1,\sigma_1^2}$, $Y \in \Phi_{\alpha_2,\sigma_2^2}$, $H_1 = \{\alpha_1 = \alpha_2\}$, les valeurs σ_1 et σ_2 étant arbitraires. Dans cet exemple, le numérateur de (22) est le même que dans l'exemple précédent, quant au dénominateur, nous l'avons calculé dans l'exemple 4 (cf. (17) ; c'était le numérateur de (8)).

Un test asymptotiquement minimax sera donc de la forme

$$\left(\frac{S_X^2 + (1-p)^2\Delta^2}{S_X^2}\right)^a \left(\frac{S_Y^2 + p^2\Delta^2}{S_Y^2}\right)^{1-a} > e^{h_\epsilon/(n_1+n_2)}\,; \qquad (23)$$

où $\Delta = \bar{x} - \bar{y}$ a pour expression

$$\Delta = (\alpha_1 - \alpha_2) + \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,\xi, \quad \xi \in \Phi_{0,1},$$
$$S_X^2/\sigma_1^2 \underset{P}{\to} 1, \quad S_Y^2/\sigma_2^2 \underset{P}{\to} 1,$$

de sorte que $\Delta \underset{P}{\to} 0$ pour l'hypothèse H_1. Cette relation est valable de toute évidence pour chaque alternative voisine. Pour trouver un test asymptotiquement équivalent à (23) de forme plus simple, considérons les parties principales des deux membres de l'inégalité (23). On a

$$\frac{a(1-p)^2\Delta^2}{S_X^2} + \frac{(1-a)p^2\Delta^2}{S_Y^2} + \Delta^4\rho_n > \frac{h_\epsilon}{n_1+n_2} + O\left(\frac{1}{(n_1+n_2)^2}\right),$$

où $\rho_n \underset{P}{\to} \rho = \text{const}$. Vu que

$$p = \frac{aS_Y^2}{aS_Y^2 + (1-a)S_X^2} + \Delta^2\rho_n', \quad \rho_n' \underset{P}{\to} \rho' = \text{const},$$

il vient

$$\frac{a(1-a)^2 S_X^2\Delta^2(n_1+n_2) + a^2(1-a)S_Y^2\Delta^2(n_1+n_2)}{(aS_Y^2 + (1-a)S_X^2)^2} +$$
$$+ \Delta^4(n_1+n_2)\rho_n'' > h_\epsilon + O\left(\frac{1}{n_1+n_2}\right),$$

où $\rho_n'' \to \rho'' = \text{const}$, $\Delta^4(n_1+n_2) \underset{P}{\to} 0$. Cette inégalité peut être mise sous la forme équivalente

$$\frac{\Delta^2(n_1+n_2)}{S_X^2/a + S_Y^2/(1-a)} > h_\epsilon + \delta_n, \quad \delta_n \underset{P}{\to} 0.$$

D'où il vient que le test

$$\frac{|\bar{x} - \bar{y}|\sqrt{n_1+n_2}}{\sqrt{S_X^2/a + S_Y^2/(1-a)}} > \sqrt{h_\epsilon} = \lambda_{\epsilon/2} \tag{24}$$

est asymptotiquement équivalent à (23) et par suite asymptotiquement minimax pour le problème de Berens-Fisher. $\lambda_{\epsilon/2}$ admet la même signification que dans l'exemple 4. Contrairement aux exemples 2, 3 et 4, la distribution exacte de la statistique du premier membre de (24) dépend pour H_1 des paramètres inconnus σ_1^2 et σ_2^2.

5. Quelques autres problèmes. Signalons encore deux classes de problèmes dont les solutions asymptotiques peuvent être acquises à l'aide du théorème 3.15.4.

1) La première classe est composée de problèmes généralisant ceux des numéros 2, 3 et 4 au cas où sont testées des hypothèses de la forme $\{\theta_1 = f(\theta_2)\}$ (par exemple $\{\theta_1 = a + b\theta_2\}$) dans les conditions du n° 2 et de la forme $\{u_1 = f(u_2)\}$ dans les conditions des n°s 3 et 4. Il est immédiat de voir que les raisonnements des n°s 2, 3 et 4 s'étendent à ce cas plus général.

2) La deuxième classe comprend des problèmes relatifs à trois échantillons et plus. Considérons par exemple le problème d'homogénéité pour trois échantillons. Soient $X \in \mathbf{P}_{\theta_1}$, $Y \in \mathbf{P}_{\theta_2}$ et $Z \in \mathbf{P}_{\theta_3}$. On teste l'hypothèse $H_1 = \{\theta_1 = \theta_2 = \theta_3\}$ contre son alternative. Supposons pour simplifier que les échantillons sont de tailles n_1, n_2 et n_3 égales à n. Considérons l'échantillon global (X, Y, Z) comme un échantillon de taille n d'éléments $(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ de densité $f_{\bar\theta}(x,y,z) = f_{\theta_1}(x) f_{\theta_2}(y) f_{\theta_3}(z)$, où $\bar\theta = (\theta_1, \theta_2, \theta_3)$. L'hypothèse H_1 sera alors équivalente au fait que $\bar\theta$ est situé sur la « courbe » $\bar\theta = g(\alpha)$, $\alpha \equiv \theta_1$, $g(\alpha) = (\alpha, \alpha, \alpha)$. Nous voyons que le problème se ramène de nouveau à celui envisagé dans le théorème 3.15.4.

§ 2. Problèmes d'homogénéité dans le cas général

1. Position du problème. Dans ce paragraphe nous étudierons deux échantillons X et Y de tailles respectives n_1 et n_2 sans postuler qu'ils sont distribués suivant une loi d'une famille paramétrique.

Le problème d'homogénéité des échantillons X et Y se présente comme suit dans le cas général. Soient $X \in \mathbf{P}_1$ et $Y \in \mathbf{P}_2$. On demande de tester l'hypothèse $H_1 = \{\mathbf{P}_1 = \mathbf{P}_2\}$ contre $H_2 = \{\mathbf{P}_1 \neq \mathbf{P}_2\}$. Ces hypothèses sont visiblement toutes deux multiples. Les distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ peuvent appartenir à une famille donnée $\mathscr{P}$ ou être arbitraires. Le principe général de construction d'un test de choix entre H_1 et H_2 reste le même que dans le chapitre 3 à une seule différence près, c'est qu'il est, comme au § 1, relatif à l'échantillon global (X, Y), de sorte que $\pi = \pi(X, Y)$ est la probabilité d'accepter H_2 pour (X, Y). Dans le cas non randomisé ($\pi = 0$ ou 1) le test π est défini par une région critique $\Omega \subset \mathscr{X}^{n_1+n_2}$ telle que H_2 est acceptée si $(X, Y) \in \Omega$. Le nombre

$$1 - \epsilon = \inf_{\mathbf{P}_1 \in \mathscr{P}} \mathbf{P}_1 \times \mathbf{P}_1((X, Y) \notin \Omega)$$

s'appelle ***niveau*** (ou ***seuil***) ***de signification*** et

$$\beta_\pi(\mathbf{P}_1, \mathbf{P}_2) = \mathbf{P}_1 \times \mathbf{P}_2((X, Y) \in \Omega), \quad \mathbf{P}_1 \in \mathscr{P}, \quad \mathbf{P}_2 \in \mathscr{P},$$

puissance du test π au « point » $(\mathbf{P}_1, \mathbf{P}_2)$.

Un test π est dit *convergent* si $\beta_\pi(\mathbf{P}_1, \mathbf{P}_2) \to 1$ lorsque $n_1 \to \infty, n_2 \to \infty$ quelles que soient $\mathbf{P}_1 \neq \mathbf{P}_2$, $\mathbf{P}_1 \in \mathscr{P}$, $\mathbf{P}_2 \in \mathscr{P}$.

Nous savons déjà que les distributions d'échantillonnage $\mathbf{P}_X^*$ et $\mathbf{P}_Y^*$ correspondant aux échantillons X et Y tendent vers $\mathbf{P}_1$ et $\mathbf{P}_2$ lorsque n_1 et n_2 augmentent. C'est pourquoi il est naturel d'utiliser, pour construire les tests d'homogénéité, des « distances » $d(\mathbf{P}_X^*, \mathbf{P}_Y^*)$ de $\mathbf{P}_X^*$ à $\mathbf{P}_Y^*$, vérifiant les conditions générales décrites dans le § 3.12. Ceci étant les tests non paramétriques et asymptotiquement non paramétriques présentent un intérêt particulier. Ces tests se définissent comme suit.

Soit $d(\mathbf{P}, \mathbf{Q})$ une distance (pas forcément une métrique) sur l'espace des distributions. Si la probabilité

$$\mathbf{P}_1 \times \mathbf{P}_1 (d\,(\mathbf{P}_X^*, \mathbf{P}_Y^*) > c) = \epsilon \tag{1}$$

est indépendante du choix de $\mathbf{P}_1$, le test π défini par les égalités

$$\pi(X, Y) = \begin{cases} 0 & \text{si } d(\mathbf{P}_X^*, \mathbf{P}_Y^*) \leqslant c, \\ 1 & \text{sinon} \end{cases} \tag{2}$$

est dit *non paramétrique*. Il est évident que le test non paramétrique construit est de niveau $1 - \epsilon$.

Les tests asymptotiquement non paramétriques se définissent de façon analogue, la relation (1) devant être valable par adjonction de l'opération $\lim\limits_{n_1 \to \infty, n_2 \to \infty}$ au premier membre. Dans ce cas le test (2) aura un niveau asymptotique égal à $1 - \epsilon$. En l'absence de non-paramétricité (exacte ou asymptotique), il est assez malaisé de construire un test d'homogénéité de niveau donné.

Considérons quelques principaux tests d'homogénéité.

2. Test de Kolmogorov-Smirnov. Supposons que $\mathbf{P}_1$ et $\mathbf{P}_2$ sont de la clacce $\mathscr{P}$ des distributions continues sur la droite et soient F_X^* et F_Y^* les fonctions de répartition empiriques respectives de $\mathbf{P}_X^*$ et $\mathbf{P}_Y^*$. Dans le test de Kolmogorov-Smirnov, la distance est

$$D_{n_1, n_2} = \sup_t |F_X^*(t) - F_Y^*(t)|.$$

Le test $D_{n_1, n_2} > c$ basé sur la statistique D_{n_1, n_2} est non paramétrique. En effet, supposons que l'hypothèse H_1 est vraie et que $F(t)$ est la fonction de répartition conjointe de X et Y. La statistique D_{n_1, n_2} peut s'écrire

$$D_{n_1, n_2} = \sup_t |G_X^*(F(t)) - G_Y^*(F(t))|, \tag{3}$$

où $G_X^*(u) = F_X^*(F^{-1}(u))$ est la fonction de répartition empirique de la loi uniforme sur $[0, 1]$ (cf. §§ 1.6, 3.12). Mais en vertu de (3) on a $D_{n_1, n_2} =$

$= \sup_u |G_X^*(u) - G_Y^*(u)|$, de sorte que la distribution de D_{n_1,n_2} est indépendante de F.

On pourrait trouver la distribution exacte de D_{n_1,n_2}. Pour $n_1 = n_2 = n$ par exemple

$$\mathbf{P}(nD_{n,n} \geqslant k) = 2(C_{2n}^n)^{-1} \sum_{j=1}^{[n/k]} (-j)^{j+1} C_{2n}^{n-jk}, \tag{4}$$

$k = 1, 2, \ldots, n$. Ce fait a été établi par Gnédenko et Koroliouk par une réduction du problème à un problème simple sur les promenades aléatoires (cf. [26]).

Au § 1.6 nous avons vu que la distribution de $n_1 G_X^*(u)$ est confondue avec celle du processus poissonnien $\zeta_1(u)$, $\zeta_1(1) = n_1$. Puisque $G_X^*(u)$ et $G_Y^*(u)$ sont indépendantes, la distribution de $G_X^*(u) - G_Y^*(u)$, $u \in [0, 1]$, est confondue avec celle d'un processus poissonnien complexe $\zeta(u)$ dans lequel les sauts de valeur $1/n_1$ et $1/n_2$ ont lieu avec les intensités respectives n_1 et n_2 ; la distribution doit être considérée sous la condition que $n_1 + n_2$ sauts se sont produits et que $\zeta(1) = 0$. Donc

$$\mathbf{P}(D_{n_1,n_2} < x) =$$
$$= \mathbf{P}\left(\sup_{u\leqslant 1} |\zeta(u)| < x/\zeta(1) = 0\,;\, n_1 + n_2 \textit{ sauts se sont produits}\right).$$

Ce fait est utilisé dans l'Annexe II pour démontrer le théorème 1.6.2 de convergence du processus $w_n(u) = \sqrt{n_1}\,(G_X^*(u) - u)$ vers un pont brownien $w^\circ(n)$ et la convergence du processus

$$w_{n_1,n_2}(u) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,(G_X^*(u) - G_Y^*(u))$$

vers un pont brownien.

Plus exactement, la distribution $f(w_{n_1,n_2})$ converge vers la distribution $f(w^\circ)$ pour une métrique uniforme quelle que soit f mesurable et continue. On en déduit immédiatement la proposition suivante dite *théorème de Smirnov*.

THÉORÈME 1.

$$\lim_{n_1\to\infty, n_2\to\infty} \mathbf{P}\left(\sqrt{\frac{n_1 n_2}{n_1 + n_2}}\, D_{n_1,n_2} < x\right) = \mathbf{P}\left(\sup_{u\leqslant 1} |w^\circ(u)| < x\right) = K(x),$$

où $K(x)$ est la fonction de Kolmogorov (cf. §§ 1.8, 3.12).

La fonction $K(x)$ étant tabulée, le théorème 1 est un outil commode pour le calcul approché du niveau de signification du test de Kolmogorov-Smirnov.

Nous proposons au lecteur de s'assurer que le test de Kolmogorov-Smirnov est convergent.

3. Test du signe. Soit $n_1 = n_2 = n$. Formons les n différences

$$\mathrm{x}_1 - \mathrm{y}_1, \ldots, \mathrm{x}_n - \mathrm{y}_n \tag{5}$$

des observations de X et Y. Si l'hypothèse H_1 est vraie et $\mathbf{P}_1 \times \mathbf{P}_1(\mathrm{x}_1 - \mathrm{y}_1 = 0) = 0$ pour tous les $\mathbf{P}_1 \in \mathscr{P}$ (de toute évidence, il en est toujours ainsi si $\mathscr{P}$ est un ensemble de distributions continues), alors

$$\mathbf{P}_1 \times \mathbf{P}_1(\mathrm{x}_1 - \mathrm{y}_1 > 0) = \mathbf{P}_1 \times \mathbf{P}_1(\mathrm{x}_1 - \mathrm{y}_1 < 0) = 1/2.$$

La statistique ν du test du signe est le nombre de différences strictement positives de (5) *). On peut construire ce test en prenant pour région critique l'ensemble

$$\Omega = \left\{(X, Y) : \left|\nu - \frac{n}{2}\right| > c\right\}.$$

Ce test est non paramétrique, puisque la distribution $\mathbf{P}_1$ est indépendante de ν et

$$\mathbf{P}_1 \times \mathbf{P}_1(\nu = k) = C_n^k 2^{-n}.$$

Le nombre c se détermine à partir de la relation

$$\sum_{k : |2k - n| \leqslant 2c} C_n^k 2^{-n} \geqslant 1 - \epsilon. \tag{6}$$

Vu que le premier membre croît de façon discrète lorsque c augmente, pour solution il faut prendre la plus petite valeur de c pour laquelle le premier membre de (6) est $\geqslant 1 - \epsilon$.

Nous voyons que l'on utilise ici un test pour éprouver l'hypothèse que la probabilité de succès dans le schéma de Bernoulli est égale à 1/2. Du point de vue du problème primitif, on teste non pas l'hypothèse d'homogénéité, mais l'hypothèse plus large que

$$\mathbf{P}_1 \times \mathbf{P}_2(\mathrm{x}_1 - \mathrm{y}_1 < 0) = \int F_1(t)\, dF_2(t) = 1/2, \tag{7}$$

où F_i est la fonction de répartition de $\mathbf{P}_i$, $i = 1, 2$. La relation (7) exprime que la médiane de la distribution de $\mathrm{x}_1 - \mathrm{y}_1$ est nulle.

*) Si certaines différences $\mathrm{x}_i - \mathrm{y}_i$ sont nulles, il faut tout simplement les abandonner et prendre pour n le nombre de différences non nulles.

Le test du signe de niveau asymptotique $1 - \epsilon$ sera de la forme

$$\pi(X, Y) = 1 \quad si \quad \frac{2\left|v - \frac{n}{2}\right|}{\sqrt{n}} > \lambda_{\epsilon/2},$$
$$\Phi_{0,1}\,[\![-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}]\!] = 1 - \epsilon. \tag{8}$$

Ce test n'est pas convergent, puisque pour $\mathbf{P}_1 \neq \mathbf{P}_2$ vérifiant (7) on a $\beta_\pi(\mathbf{P}_1, \mathbf{P}_2) \to \epsilon < 1$ lorsque $n_1 \to \infty$, $n_2 \to \infty$.

4. Test de Wilcoxon. Ce test fait largement recette pour éprouver les hypothèses d'homogénéité.

Considérons l'échantillon global (X, Y) et l'échantillon ordonné associé, c'est-à-dire l'échantillon obtenu en rangeant les éléments de (X, Y) par ordre de grandeur croissante. Nous obtenons une suite de la forme

$$\mathrm{y}^{(1)}, \mathrm{y}^{(2)}, \mathrm{x}^{(3)}, \mathrm{y}^{(4)}, \mathrm{x}^{(5)}, \ldots, \tag{9}$$

où l'indice supérieur représente le numéro de l'observation dans l'échantillon ordonné associé, et la lettre indique à quel échantillon appartient cette observation. Supposons que $r_1, r_2, \ldots, r_{n_1}$ désignent les rangs des éléments de X dans l'échantillon ordonné (9). Pour la suite (9), $r_1 = 3$ et $r_2 = 5$. On appelle *statistique de Wilcoxon* la fonction

$$U = U(X, Y) = \sum_{i=1}^{n_1} (r_i - i),$$

où $r_i - i$ est le nombre des éléments de Y inférieurs à $\mathrm{x}_{(i)}$.

Vu que les transformations monotones effectuées sur les variables ne modifient pas l'ordre des observations (9) (l'ordre sera le même pour $F_X^*(t)$, $F_Y^*(t)$ que pour $F_X^*(F^{-1}(t))$, $F_Y^*(F^{-1}(t))$, où F est la fonction de répartition), le test basé sur la statistique U sera non paramétrique.

THÉORÈME 2. *Soient $X \in \mathbf{P}_1$, $Y \in \mathbf{P}_2$ et supposons que $F_i \in \mathscr{F}$ est la fonction de répartition de $\mathbf{P}_i$, $i = 1, 2$, $\mathscr{F}$ étant la classe des fonctions de répartition continues. Supposons d'autre part que $a = n_1/(n_1 + n_2) \to a_0$ lorsque $n_1 \to \infty$ et $n_2 \to \infty$. Alors*

$$\frac{U - n_1 n_2 \mathbf{E} F_2(\mathrm{x}_1)}{\sqrt{n_1 n_2 (n_1 + n_2)}} \in \Phi_{0,\sigma^2}, \tag{10}$$

où $\sigma^2 = (1 - a_0)\, \mathbf{V} F_2(\mathrm{x}_1) + a_0\, \mathbf{V} F_1(\mathrm{y}_1)$.

Si $F_1 = F_2 = F$, alors $F_2(\mathrm{x}_1) \in \mathbf{U}_{0,1}$, $F_1(\mathrm{y}_1) \in \mathbf{U}_{0,1}$ et par suite $\mathbf{E}F_2(\mathrm{x}_1) = 1/2$, $\mathbf{V}F_2(\mathrm{x}_1) = \mathbf{V}F_1(\mathrm{y}_1) = 1/12$.

Donc, le test de Wilcoxon de niveau asymptotique $1 - \epsilon$ sera de la forme

$$\left|U - \frac{n_1 n_2}{2}\right| > \frac{\lambda_{\epsilon/2}\sqrt{n_1 n_2(n_1 + n_2)}}{2\sqrt{3}}, \tag{11}$$

$$\Phi_{0,1}\,(]-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}[) = 1 - \epsilon.$$

Sur (10) on voit que ce test vise essentiellement à éprouver l'hypothèse (comparer avec (7))

$$\int F_2(t)\,dF_1(t) = 1/2 \quad \text{ou} \quad \int (F_2(t) - F_1(t))\,dF_1(t) = 0. \tag{12}$$

Si l'on convient sans perte de généralité que $F_1(t) \equiv t$, $t \in [0, 1]$ et que l'on admette que $F_2(0) = 0$, $F_2(1) = 1$, alors en vertu de l'égalité

$$\int_0^1 (1 - F_2(t))\,dt = \mathsf{E}y_1$$

l'hypothèse testée devient $\mathsf{E}y_1 = 1/2$.

Ceci exprime que le test de Wilcoxon tout comme le test du signe est sensible essentiellement aux translations des distributions l'une par rapport à l'autre. La puissance de ces tests est assez grande (cf. exemple 1) pour de telles alternatives déplacées. Si $F_2 \neq F_1$ et que (12) soit réalisée, l'hypothèse $\{F_2 = F_1\}$ sera acceptée avec une probabilité proche de $\Phi_{0,1}\left(\left]-\frac{\lambda_{\epsilon/2}}{2\sqrt{3}\sigma}, \frac{\lambda_{\epsilon/2}}{2\sqrt{3}\sigma}\right[\right)$ en vertu du test de Wilcoxon. Donc ce test n'est pas convergent.

Démonstration du théorème 2. La statistique U peut être mise sous la forme

$$U = \sum_{i=1}^{n_1} n_2 F_Y^*(x_i) = n_1 n_2 \int F_Y^*(t)\,dF_X^*(t).$$

Posons

$$w_X(t) = \sqrt{n_1}\,(F_X^*(t) - F_1(t)), \qquad w_Y(t) = \sqrt{n_2}\,(F_Y^*(t) - F_2(t)).$$

On a alors, de toute évidence,

$$U = n_1 n_2 \int F_2(t)\,dF_1(t) + \sqrt{n_1 n_2(n_1 + n_2)} \times$$
$$\times \left[\sqrt{a}\int w_Y(t)\,dF_1(t) + \sqrt{1-a}\int F_2(t)\,dw_X(t)\right] + \sqrt{n_1 n_2}\int w_Y(t)\,dw_X(t). \tag{13}$$

Vu que $\int F_2(t)\,dw_X(t) = \int w_X(t)\,dF_2(t)$, donc que la deuxième et la troisième intégrale de (13) sont de la même forme et sont indépendantes, pour prouver le théorème il suffit de s'assurer que

$$\int w_Y(t)\,dF_1(t) \Subset \Phi_{0,\sigma_1^2}, \quad \sigma_1^2 = \mathsf{V}F_1(y_1), \tag{14}$$

et que
$$\frac{1}{\sqrt{n_1 + n_2}} \int w_Y(t)\, dw_X(t) \underset{P}{\to} 0. \tag{15}$$

Le théorème 1.6.2 nous donne
$$\int w_Y(t)\, dF_1(t) \in \int w^\circ(F_2(t))\, dF_1(t), \tag{16}$$

où $w^\circ(u)$ est un pont brownien. Pour déterminer la distribution de la dernière intégrale, on remarquera que les trajectoires du processus wienérien $w(u)$ sont presque sûrement continues [11], $w^\circ(u) = w(u) - uw(1)$, et que par conséquent l'intégrale (16) est par définition la quantité vers laquelle convergent presque sûrement pour $N \to \infty$ les sommes
$$\sum_{i=1}^{N} w(F_2(t_i))\, \Delta_i F_1 - m_1 w(1), \tag{17}$$

où $m_1 = \int F_2(t)\, dF_1(t)$ et $\{t_i\}_{i=0}^{N}$ est une subdivision de l'axe réel, $\Delta_i g = g(t_i) - g(t_{i-1})$,

$w(F_2(t_i)) = \sum_{l=1}^{i} \Delta_l w(F_2)$, $w(1) = \sum_{l=1}^{N} \Delta_l w(F_2)$. D'après la transformation d'Abel
$$\sum_{i=1}^{N} \left(\sum_{l=1}^{i} d_l \right) b_i = \sum_{l=1}^{N} \left(\sum_{i=l}^{N} b_i \right) d_l .$$

Donc, (17) est égale à
$$\sum_{l=1}^{N} (1 - F_1(t_{l-1}) - m_1)\, \Delta_l w(F_2). \tag{18}$$

Ici $1 - m_1 = \int F_1(t)\, dF_2(t) = m_2$ et $\Delta_l w(F_2)$ sont des variables aléatoires normales indépendantes de paramètres $(0, \Delta_l F_2)$. La distribution (17), (18) sera donc normale, de moyenne nulle et de variance
$$\sum_{l=1}^{N} (m_2 - F_1(t_{l-1}))^2 \Delta_l F_2 \to \int (m_2 - F_1(t))^2\, dF_2(t) = \mathbf{V}F_1(y_1).$$

Ceci prouve (14).

Pour établir (15) *), le plus simple est d'estimer la variance de l'intégrale de (15). En approchant encore l'intégrale par une somme finie, on s'assure que la variance
$$\mathbf{V}_{X,Y} = \mathbf{E}\left(\int w_Y(t)\, dw_X(t) \right)^2$$

est bornée lorsque $n_1 \to \infty$, $n_2 \to \infty$. Ceci et l'inégalité de Tchébychev entraînent (15). Nous glissons sur la démonstration du fait que $\mathbf{V}_{X,Y}$ est bornée à cause de la lourdeur et de la routine des calculs. ◄

Voir également [35] au sujet des tests du signe et de Wilcoxon._

---

*) L'intégrale de (15) converge en loi vers $\int w^\circ(F_2(t))\, dw^\circ(F_1(t))$.

EXEMPLE 1. Nous avons déjà signalé que les tests du signe et de Wilcoxon étaient les plus sensibles aux translations. Il serait intéressant de comparer leurs puissances à celle du test optimal d'homogénéité pour une famille $\mathscr{P}$ de distributions se déduisant l'une de l'autre par une translation. Plus exactement, soit

$$\mathscr{P} = \{\Phi_{\alpha,1}\}, \quad \mathbf{P}_1 = \Phi_{\alpha_1,1}, \quad \mathbf{P}_2 = \Phi_{\alpha_2,1}, \quad n_1 = n_2 = n.$$

Dans ce cas le théorème 1.1 affirme l'existence d'un test asymptotiquement minimax π_0 de niveau $1 - \epsilon$ de l'hypothèse $H_1 = \{\mathbf{P}_1 = \mathbf{P}_2\} = \{\alpha_1 = \alpha_2\}$ contre $H_2^b = \{|\alpha_1 - \alpha_2| \geqslant b/\sqrt{n}\}$ qui est de la forme

$$|\bar{x} - \bar{y}| > \lambda_{\epsilon/2}\sqrt{2/n}, \qquad \Phi_{0,1}(]-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}[) = 1 - \epsilon$$

(le lecteur peut s'assurer seul que l'inégalité de cet exemple est équivalente à (1.3), (1.4)). Utilisons ce test comme un étalon de comparaison avec d'autres tests et considérons l'alternative $(\mathbf{P}_1, \mathbf{P}_2)$, où $\alpha_2 = \alpha_1 + c/\sqrt{n}$ (nous considérons des alternatives voisines pour éviter d'avoir affaire au problème des grands écarts). Il est évident que dans ce cas $(\bar{x} - \bar{y}) \in$ $\in \Phi_{-c/\sqrt{n},2/n}$. Donc

$$\beta_{\pi_0}(\mathbf{P}_1, \mathbf{P}_2) = \mathbf{P}_1 \times \mathbf{P}_2(|\bar{x} - \bar{y}| > \lambda_{\epsilon/2}\sqrt{2/n}) =$$
$$= 1 - \Phi_{-c/\sqrt{2},1}(]-\lambda_{\epsilon/2}, \lambda_{\epsilon/2}[) =$$
$$= 1 - \Phi_{0,1}(]-\lambda_{\epsilon/2} + c/\sqrt{2}, \lambda_{\epsilon/2} + c/\sqrt{2}[) \equiv \beta_0(c). \quad (19)$$

Considérons maintenant le test du signe (8) et désignons-le par π_1. En développant en série suivant les puissances de $c/\sqrt{n}$, on obtient ($\Phi_{\alpha,\sigma^2}(x) =$ $= \Phi_{\alpha,\sigma^2}(]-\infty, x[)$)

$$\mathbf{P}_1 \times \mathbf{P}_2(x_1 - y_1 < 0) = \Phi_{0,2}\left(\frac{c}{\sqrt{n}}\right) = \frac{1}{2} + \frac{c}{\sqrt{n}} \cdot \frac{1}{2\sqrt{\pi}} + O\left(\frac{1}{n}\right).$$

Donc, au point $(\mathbf{P}_1, \mathbf{P}_2)$

$$\frac{2}{\sqrt{n}}\left(\nu - \frac{n}{2} + \frac{c\sqrt{n}}{2\sqrt{\pi}}\right) \in \Phi_{0,1}.$$

Pour le test π_1 de niveau asymptotique $1 - \epsilon$ on a par conséquent

$$\beta_{\pi_1}(\mathbf{P}_1, \mathbf{P}_2) = \mathbf{P}_1 \times \mathbf{P}_2\left(2\left|\nu - \frac{n}{2}\right| > \lambda_{\epsilon/2}\sqrt{n}\right) \to$$
$$\to 1 - \Phi_{0,1}\left(\left]-\lambda_{\epsilon/2} + \frac{c}{\sqrt{\pi}}, \lambda_{\epsilon/2} + \frac{c}{\sqrt{\pi}}\right[\right).$$

Considérons enfin le test π_2 de Wilcoxon (cf. (11)) qui est ici de la forme

$$\left| U - \frac{n^2}{2} \right| > \frac{\lambda_{\epsilon/2} n^{3/2}}{\sqrt{6}}.$$

Il est évident que la statistique U est invariante par une translation des éléments de X et Y. On peut donc admettre que $\mathbf{P}_1 = \Phi_{0,1}$, $\mathbf{P}_2 = \Phi_{c/\sqrt{n},1}$ et par suite

$$\mathbf{E}F_2(\mathrm{x}_1) = \int F_2(t)\, dF_1(t) = \int \Phi_{0,1}\left(t - \frac{c}{\sqrt{n}}\right) d\Phi_{0,1}(t) =$$

$$= \Phi_{0,2}\left(-\frac{c}{\sqrt{n}}\right) = \frac{1}{2} - \frac{c}{\sqrt{n}} \cdot \frac{1}{2\sqrt{\pi}} + O\left(\frac{1}{n}\right).$$

Comme $\mathbf{V}F_2(\mathrm{x}_1) \to \mathbf{V}F_1(\mathrm{x}_1) = 1/12$, $\mathbf{V}F_1(\mathrm{y}_1) \to \mathbf{V}F_1(\mathrm{x}_1) = 1/12$, il vient en vertu du théorème 2

$$\beta_{\pi_2}(\mathbf{P}_1, \mathbf{P}_2) = \mathbf{P}_1 \times \mathbf{P}_2\left(\left| U - \frac{n^2}{2} \right| > \frac{\lambda_{\epsilon/2} n^{3/2}}{\sqrt{6}}\right) =$$

$$= 1 - \mathbf{P}_1 \times \mathbf{P}_2\left(-\lambda_{\epsilon/2} + c\sqrt{\frac{3}{2\pi}} \leqslant \right.$$

$$\left. \leqslant \sqrt{6}\, n^{-3/2}\left(U - \frac{n^2}{2} + \frac{n^{3/2} c}{2\sqrt{\pi}}\right) \leqslant \lambda_{\epsilon/2} + c\sqrt{\frac{3}{2\pi}}\right) =$$

$$= 1 - \Phi_{0,1}\left(\left]-\lambda_{\epsilon/2} + c\sqrt{\frac{3}{2\pi}}, \lambda_{\epsilon/2} + c\sqrt{\frac{3}{2\pi}}\right[\right).$$

Remarquons maintenant que $\beta_0(c)$ (cf. 19)) est une fonction monotone strictement croissante de c et que pour les grands n

$$\beta_{\pi_1}(\mathbf{P}_1, \mathbf{P}_2) \approx \beta_0\left(\sqrt{\frac{2}{\pi}}\, c\right), \quad \beta_{\pi_2}(\mathbf{P}_1, \mathbf{P}_2) \approx \beta_0\left(\sqrt{\frac{3}{\pi}}\, c\right).$$

Donc, pour tout $c > 0$ le plus puissant des tests π_0, π_1 et π_2 est, comme on s'y attendait, le test π_0. Viennent ensuite le test de Wilcoxon et le test du signe ; à noter que le test de Wilcoxon le cède de très peu au test π_0, puisque $\sqrt{3/\pi} \approx 0{,}977$.

Si l'on considère des échantillons X' et Y' de taille $n' > n$ pour le même biais $\alpha_2 - \alpha_1 = c/\sqrt{n}$, alors pour déterminer la puissance des tests $\pi_j(X', Y')$ au point $(\mathbf{P}_1, \mathbf{P}_2)$ à l'aide de la même procédure de calcul, il faut envisager le problème précédent pour une nouvelle valeur de c égale à $c' = c\sqrt{n'}/\sqrt{n}$ ($\alpha_2 - \alpha_1$ sera alors égale à $c'/\sqrt{n'}$). Donc, au point $(\mathbf{P}_1, \mathbf{P}_2)$ les puissances de $\pi_1(X', Y')$ et de $\pi_2(X', Y')$ seront approximativement

égales à

$$\beta_0\left(\sqrt{\frac{2}{\pi}}c'\right)=\beta_0\left(\sqrt{\frac{2n'}{\pi n}}c\right),\quad \beta_0\left(\sqrt{\frac{3}{\pi}}c'\right)=\beta_0\left(\sqrt{\frac{3n'}{\pi n}}c\right).$$

L'identification $\frac{2n'}{\pi n} = 1$, $\frac{3n'}{\pi n} = 1$ nous donne les valeurs $n' = \frac{\pi}{2}n$, $n' = \frac{\pi}{3}n$ (indépendantes de c) pour le nombre d'observations qu'il faut effectuer pour obtenir la même puissance avec les tests π_1 et π_2 qu'avec le test π_0 pour n observations. Par exemple, pour obtenir les mêmes résultats il faut effectuer 100 observations pour le test π_0, ≈ 105 pour le test π_2 et ≈ 157 pour le test π_1.

On obtiendrait des résultats foncièrement différents si l'on testait l'homogénéité pour la famille $\mathscr{P} = \{\Phi_{0,\sigma^2}\}$. Les tests du signe et de Wilcoxon seraient non convergents dans ce cas. Plus, le test du signe de niveau $1 - \epsilon$ serait en fait identique au test $\pi \equiv \epsilon$ qui est indépendant des échantillons, puisque $\mathbf{E}(x_1 - y_1) = 0$ et $\mathbf{P}_1 \times \mathbf{P}_2(x_1 - y_1 > 0) = 1/2$ pour tout couple de distributions $\mathbf{P}_1$ et $\mathbf{P}_2$ de $\mathscr{P}$. Pour ce problème on pourrait envisager d'autres tests non paramétriques basés sur les statistiques r_i, par exemple le test $\sum_{i=0}^{n_1} (r_{i+1} - r_i)^2, r_0 = 0, r_{n_1+1} = n_2$, qui rappelle par ses propriétés le test de Moran (§ 3.12).

5. Le test du χ^2 comme test asymptotiquement optimal de l'homogénéité au vu de données groupées. Dans ce numéro nous admettrons que les données sont groupées dans les deux échantillons X et Y de tailles respectives n_1 et n_2 (cf. § 3.16). Au lieu des échantillons X et Y on peut utiliser dans ce cas les vecteurs $\nu = (\nu_1, \ldots, \nu_r)$ et $\mu = (\mu_1, \ldots, \mu_r)$ des fréquences des observations respectivement des échantillons X et Y contenues dans les intervalles de groupement $\Delta_1, \ldots, \Delta_r$. Désignons par $\theta_i = (\theta_{i1}, \ldots, \theta_{ir})$, $i = 1, 2$, les vecteurs des probabilités d'accès des observations respectives de X et de Y aux intervalles $\Delta_1, \ldots, \Delta_r$, de sorte que $\theta_{1l} = \mathbf{P}(x_j \in \Delta_l)$, $\theta_{2l} = \mathbf{P}(y_j \in \Delta_l)$. Les échantillons grossis X et Y peuvent alors être traités comme des échantillons distribués suivant des lois des familles paramétriques $\{\mathbf{B}_{\theta_1}\}$ et $\{\mathbf{B}_{\theta_2}\}$ respectivement. Le problème devient donc paramétrique et l'on peut utiliser les résultats développés dans l'exemple 1 du paragraphe précédent. Il résulte de cet exemple que si nous testons l'hypothèse d'homogénéité $H_1 = \{\theta_1 = \theta_2\}$ dans le cas où le paramètre θ est localisé, c'est-à-dire dans le cas où les valeurs θ_1 et θ_2 sont situées au voisinage du point $\theta_0 = (\theta_{01}, \ldots, \theta_{0r})$, un test asymptotiquement minimax de niveau

asymptotique $1 - \epsilon$ de H_1 contre

$$H_2^b = \left\{ \sum_{i=1}^{r} \frac{(\theta_{1i} - \theta_{2i})^2}{\theta_{0i}} \geqslant \frac{b^2}{n_2} \right\}$$

sera de la forme

$$\sum_{i=1}^{r} \left(\frac{\nu_i}{n_1} - \frac{\mu_i}{n_2} \right)^2 \frac{n_1 n_2}{\nu_i + \mu_i} \geqslant h_\epsilon ,$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ d'une distribution du χ^2 à $r - 1$ degrés de liberté. Ceci n'est autre qu'un test du χ^2 d'homogénéité au vu de données groupées.

Pour test asymptotiquement équivalent on pourrait envisager le test

$$\sum_{i=1}^{r} \nu_i \ln \frac{\nu_i}{n_1} + \sum_{i=1}^{r} \mu_i \ln \frac{\mu_i}{n_2} - \sum_{i=1}^{r} (\nu_i + \mu_i) \ln \frac{\nu_i + \mu_i}{n_1 + n_2} > \frac{h_\epsilon}{2} .$$

§ 3. Problèmes de régression

1. Position du problème. Dans les applications on est souvent confronté à des problèmes portant sur des observations dont les distributions varient dans les expériences en fonction de certains paramètres caractérisant ces expériences. Désignons par

$$\mathrm{x}_i = (\mathrm{x}_{i,1}, \ldots, \mathrm{x}_{i,r})$$

l'ensemble des valeurs de ces paramètres durant la i-ième expérience, $i = 1, \ldots, n$. Les valeurs $\mathrm{x}_{i,k}$ sont définies soit par l'expérimentateur, soit par la nature du phénomène étudié. Désignons le vecteur $(\mathrm{x}_{1,k}, \ldots, \mathrm{x}_{n,k})$ par X_k et la matrice $\begin{pmatrix} X_1 \\ \vdots \\ X_r \end{pmatrix} = (\mathrm{x}_1^T, \ldots, \mathrm{x}_n^T)$ par X. Donc, contrairement à ce qui précède, X est une $(r \times n)$-matrice dont les éléments peuvent être des nombres non aléatoires quelconques dont la nature nous sera indifférente. Nous désignerons le vecteur des observations par $Y = (\mathrm{y}_1, \ldots, \mathrm{y}_n)$.

Les problèmes de régression sont basés sur l'hypothèse que les observations y_i sont de la forme

$$\mathrm{y}_i = \alpha_1 \mathrm{x}_{i,1} + \ldots + \alpha_r \mathrm{x}_{i,r} + \xi_i , \quad i = 1, \ldots, n, \tag{1}$$

où $\alpha = (\alpha_1, \ldots, \alpha_r)$ sont des constantes inconnues, $\xi_i \in \Phi_{0,\sigma^2}$ sont indépendantes.

La constante α_1 joue souvent un rôle particulier, car dans bien des cas elle met un terme constant en évidence dans la représentation (1), ce qui correspond au fait que dans la matrice X on admet *a priori* que $X_1 = (1, ..., 1)$ $(x_{i,1} \equiv 1)$. Cette hypothèse ne sera pas utilisée dans la suite. Les variables aléatoires ξ_i figurent des bruits, des fluctuations ou des erreurs de mesure.

Le relation (1) peut être mise sous la forme matricielle

$$Y = \alpha X + \xi. \tag{2}$$

Une régression de la forme (1), (2) est dite linéaire (aussi bien en α qu'en X). Sont problèmes de régression aussi bien le problème d'estimation des paramètres inconnus α et σ^2 sachant que (1), (2) sont vraies, que le problème de test de l'hypothèse que (1), (2) sont valables. Dans les deux cas, on part de l'échantillon (X, Y). Le terme d'« échantillon » est pris ici dans une acception plus large qu'auparavant et représente un ensemble d'observations qui ne sont pas nécessairement de la même nature. Rappelons par ailleurs que le premier des deux « échantillons » X et Y peut être non aléatoire. La matrice X est parfois appelée *regresseur* et le vecteur Y, *réponse*.

Le modèle de régression (1), (2) est très général du point de vue de la forme de la dépendance de y_i par rapport aux paramètres. Si l'on admet par exemple que $x_{i,k} = \psi_k(z_i)$, où $\psi_1, ..., \psi_r$ est un ensemble donné de fonctions et z_i les valeurs d'un paramètre scalaire, on obtient le modèle

$$y_i = \alpha_1\psi_1(z_i) + ... + \alpha_r\psi_r(z_i) + \xi_i, \quad i = 1, ..., n, \tag{3}$$

de régression en les fonctions arbitraires $\psi_1, ..., \psi_r$ (qui est encore linéaire en α). Si $\psi_1(z) \equiv 1$, $\psi_2(z) \equiv z$ et $r = 2$, on obtient un modèle de régression *linéaire élémentaire* (de dimension un) (cf. fig. 6).

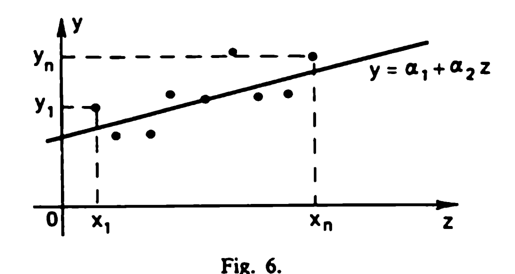


Fig. 6.

Le modèle général (1), (2) est parfois appelé modèle de régression *ensembliste* pour faire la distinction avec le modèle élémentaire. Nous verrons qu'en général les problèmes de régression sont reliés à l'étude (à l'existence) d'une dépendance fonctionnelle $y = \varphi(x)$ pour une classe donnée de

fonctions φ dans les cas où les observations de la variable y pour x donnée sont affectées d'écarts aléatoires.

Les lignes $X_1, \ldots, X_r$ de la matrice X de (2) sont généralement choisies *linéairement indépendantes* (sinon il serait impossible d'estimer les coordonnées de α). Nous adopterons cette convention qui exprime que la matrice X est de rang r.

Il apparaît plus commode d'avoir parfois affaire à des vecteurs $X_1, \ldots$ $\ldots, X_r$ orthogonaux, c'est-à-dire vérifiant la condition $(X_i, X_j) = 0, i \neq j$, où (a, b) désigne le produit scalaire. Si l'ensemble des vecteurs linéairement indépendants $\{X_k\}$ ne possède pas cette propriété, on peut l'orthogonaliser en introduisant de nouveaux vecteurs

$$\begin{aligned} X_1' &= X_1, \\ X_2' &= X_2 + a_{2,1}X_1, \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ X_r' &= X_r + a_{r,r-1}X_{r-1} + \ldots + a_{r,1}X_1. \end{aligned} \tag{4}$$

Les coefficients $a_{k,j}$ se déterminent facilement à partir des conditions d'orthogonalité $X_k' \perp X_j'$, $k \neq j$, de sorte que, par exemple, $a_{2,1} = -\dfrac{(X_1, X_2)}{(X_1, X_1)}$. Les relations (4) peuvent être mises sous la forme $X' = AX$, où A est une matrice trigonale inversible (dont la diagonale principale est composée d'unités). On en déduit que $X = A^{-1}X'$, $Y = \alpha A^{-1}X' + \xi$. On est conduit à un problème de régression de coefficients $\beta = \alpha A^{-1}$. Le vecteur α est manifestement restitué par β à l'aide de l'égalité $\alpha = \beta A$.

Pour une régression linéaire élémentaire, la condition d'orthogonalité de $X_1 = (1, \ldots, 1)$ et $X_2 = (z_1, \ldots, z_n)$ équivaut à la condition $\sum z_i = 0$ qui peut être visiblement satisfaite par un changement d'origine de la variable z.

2. Estimation des paramètres. On admettra dans la suite que $r < n$ et que les vecteurs X_k, $k = 1, \ldots, r$, sont linéairement indépendants. Dans le cas de la régression (1), (2), la fonction de vraisemblance de l'observation Y pour X donné est égale à

$$f_{\alpha,\sigma^2}(Y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{r}\alpha_k x_{i,k}\right)^2\right\} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{|Y-\alpha X|^2}{2\sigma^2}\right\}. \tag{5}$$

La fonction (5) dépend du paramètre $\theta = (\alpha, \sigma^2)$. A noter que si l'on traite (5) comme la fonction de vraisemblance non pas d'une seule observation Y (ou (X, Y)) mais de n observations $y_1, \ldots, y_n$, elle ne correspondra pas à un échantillon distribué suivant une loi d'une seule famille paramétrique. Les observations y_i suivent des lois différentes $\Phi_{\gamma_i, \sigma^2}$, $\gamma_i = \sum_{k=1}^{r} \alpha_k x_{i,k}$, dépendant de x_i. Donc, les considérations des chapitres précédents dans lesquelles les éléments de l'échantillon suivaient la même distribution ne passent pas ici.

Ainsi, nous traiterons (5) comme une fonction de vraisemblance de l'observation (X, Y). Appliquons la méthode du maximum de vraisemblance. On voit directement sur (5) que l'estimation du maximum de vraisemblance $\alpha^* = \hat{\alpha}^*$ qui maximise $f_\theta(Y)$ par rapport à α est une estimation qui minimise $|Y - \alpha X|^2$. Donc, la méthode du maximum de vraisemblance coïncide ici avec la « méthode des moindres carrés ».

Designons par $\mathscr{L}[X]$ le sous-ensemble engendré par les vecteurs $X_1, \ldots$ $\ldots, X_r$. Ce sous-espace est composé des points de la forme αX où α parcourt les valeurs de R^r, il est de dimension r et contient le point $\beta = \alpha^* X$ le moins éloigné de Y (fig. 7). La valeur de β est définie de façon unique par la

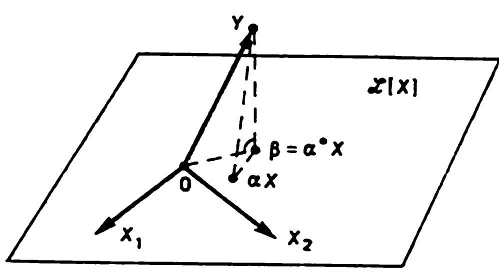


Fig. 7.

condition d'orthogonalité de $Y - \beta$ à $\mathscr{L}[X]$ ou ce qui est équivalent par les r conditions

$$(Y - \alpha^* X, X_k) = (Y - \alpha^* X)\, X_k^T = 0, \quad k = 1, \ldots, r.$$

Ces conditions peuvent s'écrire sous la forme matricielle $(Y - \alpha^* X)X^T =$ $= 0$. D'où il vient

$$\alpha^* = YX^T(XX^T)^{-1}. \tag{6}$$

La matrice inverse $(XX^T)^{-1}$ (d'ordre r) existe, puisque la matrice $D =$ $= XX^T$ est définie positive. En effet, on a vu qu'il existe une matrice non dégénérée A telle que les lignes de la matrice $X' = AX$ sont orthogonales.

Donc, la matrice D peut être mise sous la forme

$$XX^T = A^{-1}X'(X')^T(A^{-1})^T = A^{-1}B(A^{-1})^T,$$

où $B = X'(X')^T$ est une matrice diagonale d'éléments

$$(X_i', X_j') = \begin{cases} |X_i'|^2 > 0 & \text{si} \quad i = j; \\ 0 & \text{si} \quad i \neq j. \end{cases}$$

Par conséquent B est définie positive et $aBa^T > 0$ pour tout $a \in R^r$, $a \neq 0$. En posant $b = aA$, on obtient $bDb^T = aAXX^TA^Ta^T = aBa^T > 0$ pour tout $b \in R^r$, $b \neq 0$, ce que nous voulions.

Si X_k sont orthogonaux, on déduit de (6) que $\alpha_k^* = \dfrac{(Y, X_k)}{(X_k, X_k)}$.

On aurait pu établir (6) d'une autre façon : en dérivant (5) par rapport à α_k et en égalant les dérivées à zéro.

La différence $Y - \alpha^*X$ est parfois appelée *résidu*. Ce résidu est orthogonal à $\mathscr{L}[X]$, donc à tout vecteur $\gamma X \in \mathscr{L}[X]$, $\gamma \in R^r$. Si l'on prend $\gamma = \alpha^* - \alpha$, on déduit de l'égalité $Y - \alpha X = Y - \alpha^*X + (\alpha^* - \alpha)X$ que

$$|Y - \alpha X|^2 = |Y - \alpha^*X|^2 + |(\alpha^* - \alpha)X|^2. \tag{7}$$

Trouvons maintenant un estimateur du maximum de vraisemblance pour σ^2. On voit sur (5) que ce sera le même estimateur que pour une famille normale (on peut de nouveau dériver (5) par rapport à σ et égaler la dérivée à zéro), de sorte que

$$(\hat{\sigma}^2)^* = \frac{1}{n}|Y - \alpha^*X|^2. \tag{8}$$

Posons

$$(\sigma^2)^* = \frac{1}{n-r}|Y - \alpha^*X|^2 = \frac{n}{n-r}(\hat{\sigma}^2)^*. \tag{9}$$

E_l désignera dans la suite la matrice unité d'ordre l, $\sigma^* = \sqrt{(\sigma^2)^*}$.

THÉORÈME 1. *Les estimateurs* (6) *et* (9) *sont des estimateurs efficaces sans biais indépendants pour les paramètres α et σ^2. De plus*

$$(\alpha^* - \alpha)D^{1/2} \in \Phi_{0,\sigma^2E_r}, \quad D = XX^T, \tag{10}$$

$$(n - r)(\sigma^2)^*/\sigma^2 = |Y - \alpha^*X|^2/\sigma^2 \in \mathrm{H}_{n-r}. \tag{11}$$

Si X_k sont orthogonaux, α_k^* sont indépendants et

$$(\alpha_k^* - \alpha_k)|X_k| \in \Phi_{0,\sigma^2}. \tag{12}$$

COROLLAIRE 1. *De* (10) *et* (11) *on déduit que*

$$\frac{(\alpha^* - \alpha)D(\alpha^* - \alpha)^T}{(n - r)(\sigma^2)^*} = \frac{|(\alpha^* - \alpha)X|^2}{|Y - \alpha^* X|^2} \in \mathbf{F}_{r,n-r}. \tag{13}$$

Soient $\bar{\alpha}$ *et* $\bar{\alpha}^*$ *les sous-vecteurs de dimension* $l \leqslant r$ *des vecteurs* α *et* α^* *composés par les coordonnées d'indices fixes* $k_1, \ldots, k_l$ *et soit* $\bar{X}$ *la matrice formée par les lignes* $X_{k_1}, \ldots, X_{k_l}$. *Si* X_k, $k = 1, \ldots, r$, *sont orthogonaux, alors*

$$(\bar{\alpha}^* - \bar{\alpha})(\bar{X}\bar{X}^T)^{1/2} \in \mathbf{\Phi}_{0,\sigma^2 E_l}, \quad (\alpha_k^* - \alpha_k)|X_k|/\sigma^* \in \mathbf{T}_{n-r}. \tag{14}$$

DÉMONSTRATION du théorème 1. Comme $YX^T = \alpha XX^T + \xi X^T$, il vient

$$\alpha = (YX^T - \xi X^T)D^{-1}, \quad \alpha^* - \alpha = \xi X^T D^{-1}. \tag{15}$$

La matrice des moments d'ordre deux du vecteur $(\alpha^* - \alpha)D^{1/2}$ est égale à

$$\mathsf{E}D^{1/2}(\alpha^* - \alpha)^T(\alpha^* - \alpha)D^{1/2} = D^{1/2}D^{-1}X\,\mathsf{E}\,\xi^T\xi X^T D^{-1}D^{1/2} = \sigma^2 E_r.$$

Les composantes de ce vecteur sont indépendantes, car normales, et $\frac{1}{\sigma^2}|(\alpha^* - \alpha)D^{1/2}|^2 \in \mathbf{H}_r$. En vertu de (7) et (9), il vient d'autre part

$$(n - r)(\sigma^2)^* = |Y - \alpha^* X|^2 = |\xi|^2 - |(\alpha^* - \alpha)X|^2.$$

Assurons-nous maintenant que les vecteurs α^* et $Y - \alpha^* X$ (donc α^* et σ^*) sont indépendants. Ces vecteurs étant normaux, il suffit de vérifier que les coefficients des corrélations entre leurs composantes sont nuls ou, ce qui est équivalent, que la matrice des covariances $\mathsf{E}(\alpha^* - \alpha)^T(Y - \alpha^* X)$ est nulle. Remarquons qu'en vertu de (6)

$$\alpha^* X = YX^T(XX^T)^{-1}X = YX^T D^{-1}X$$

et le vecteur $\alpha^* X$ est le projeté de Y sur $\mathscr{L}[X]$. Le projecteur associé à la matrice $\Pi = X^T D^{-1}X$ est doué des propriétés évidentes : $\Pi^2 = \Pi$, $BX\Pi = BX$ pour toute matrice B à r colonnes. Donc, en vertu de (15)

$$\mathsf{E}(\alpha^* - \alpha)^T(Y - \alpha^* X) = \mathsf{E}D^{-1}X\xi^T(\xi - \xi X^T D^{-1}X) = D^{-1}X\sigma^2(E_n - \Pi) = 0.$$

Prouvons maintenant (11). La relation (7) entraîne

$$|Y - \alpha^* X|^2 = |\xi|^2 - |(\alpha^* - \alpha)X|^2 = |\xi|^2 - |(\alpha^* - \alpha)D^{1/2}|^2,$$

où $\frac{1}{\sigma^2}|\xi|^2 \in \mathbf{H}_n$, $\frac{1}{\sigma^2}|(\alpha^* - \alpha)D^{1/2}|^2 \in \mathbf{H}_r$ (cf. (10)). La proposition (11) découle de ces relations et du lemme (1).

LEMME 1. *Si $\eta = \eta_1 + \eta_2$, où η_1 et η_2 sont indépendantes, $\eta \in \mathbf{H}_n$, $\eta_1 \in \mathbf{H}_r$, alors $\eta_2 \in \mathbf{H}_{n-r}$.*

DÉMONSTRATION. Si l'on désigne par $\varphi(t)$ la fonction caractéristique de la distribution H_1 : $\varphi(t) = (1 + 2it)^{-1/2}$, alors

$$\mathbf{E}e^{i\eta t} = \varphi(t)^n = \varphi(t)^r \cdot \mathbf{E}e^{i\eta_2 t}.$$

Comme $\varphi(t) \neq 0$ sur la droite réelle, il vient $\mathbf{E}e^{i\eta_2 t} = \varphi(t)^{n-r}$. ◄

Que les estimateurs α^*, $(\sigma^2)^*$ soient sans biais résulte manifestement de (10) et (11) ($\mathbf{E}\eta = l$ si $\eta \in \mathbf{H}_l$).

Reste à prouver que $\theta^* = (\alpha^*, (\sigma^2)^*)$ est efficace. Remarquons à cet effet que la famille paramétrique (5) est de type exponentiel, puisque (5) se représente sous la forme (cf. (2.15.1))

$$f_\theta(Y) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}(|Y|^2 - 2(Y, \alpha X) + |\alpha X|^2)\right\} =$$

$$= h(Y) \exp\left\{\sum_{k=1}^{r+1} a_k(\theta)\, U_k(Y) + V(\theta)\right\},$$

où

$$h(Y) = (2\pi)^{-n/2}, \qquad V(\theta) = -n \ln \sigma - \frac{1}{2\sigma^2}|\alpha X|^2,$$

$$a_k(\theta) = \frac{\alpha_k}{\sigma^2}, \qquad U_k(Y) = (Y, X_k), \qquad k = 1, \ldots, r,$$

$$a_{r+1}(\theta) = -\frac{1}{2\sigma^2}, \qquad U_{r+1}(Y) = |Y|^2.$$

Les conditions des théorèmes 2.15.1 et 2.15.2 étant satisfaites, la statistique $U = (U_1(X), \ldots, U_{r+1}(X))$ (et avec elle θ^*) est une statistique exhaustive complète minimale. D'où l'efficacité de θ^* (cf. corollaire 2.15.1).

La proposition (12) découle visiblement de (10), puisque pour les X_k orthogonaux, la matrice $D^{1/2}$ est diagonale d'éléments $|X_k|$. ◄

REMARQUE 1. Hotelling (cf. [73]) a prouvé que $\mathbf{V}\alpha_k^* \geqslant \sigma^2/|X_k|^2$, l'égalité n'étant réalisée que si les X_k sont orthogonaux. Si donc l'on envisage de

réaliser une expérience pour des valeurs données de $|X_k|$, la façon optimale de choisir X est de rendre les X_k orthogonaux.

REMARQUE 2. Il serait intéressant de comparer la matrice des moments d'ordre deux de l'estimateur θ^* avec la borne inférieure des estimateurs sans biais, définie, en vertu de l'inégalité multidimensionnelle de Rao-Cramer, par la matrice $I^{-1}(\theta)$, où $I(\theta)$ est la matrice d'information de Fisher

$$I(\theta) = \| I_{ij}(\theta) \|, \quad I_{ij}(\theta) = \mathsf{E}_\theta \frac{\partial L}{\partial \theta_i} \cdot \frac{\partial L}{\partial \theta_j}, \quad L = L(Y;\theta) = \ln f_\theta(Y).$$

Ici $\theta_k = \alpha_k$, $k = 1, \ldots, r$, $\theta_{r+1} = \sigma^2$. Supposons pour simplifier que les X_k sont orthogonaux. L'indépendance des θ_k^* entraîne que la matrice $\mathsf{E}_\theta(\theta^* - \theta)^T(\theta^* - \theta)$ sera diagonale d'éléments

$$\mathsf{E}_\theta(\alpha_k^* - \alpha)^2 = \frac{\sigma^2}{|X_k|^2}, \qquad k = 1, \ldots, r,$$

$$\mathsf{E}_\theta\,((\sigma^2)^* - \sigma^2)^2 = \mathsf{E}\left(\frac{\sigma^2 \chi^2_{n-r}}{n-r} - \sigma^2\right)^2 = \frac{\sigma^4}{n-r}\,\mathsf{E}\,(\chi_1^2 - 1)^2 = \frac{2\sigma^4}{n-r},$$

où $\chi_l^2 \in \mathbf{H}_l$.

D'autre part, vu que

$$\frac{\partial L}{\partial \alpha_k} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} \alpha_j x_{ij}\right) x_{ik} = \frac{1}{\sigma^2}(Y - \alpha X) X_k^T,$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{r} \alpha_j x_{ij}\right)^2 = \frac{1}{2\sigma^2}\left(\frac{|Y - \alpha X|^2}{\sigma^2} - n\right),$$

on trouve pour la matrice $I(\theta)$, $k = , \ldots, r$,

$$I_{kk}(\theta) = \mathsf{E}_\theta \frac{1}{\sigma^4} X_k (Y - \alpha X)^T (Y - \alpha X) X_k^T =$$

$$= \frac{1}{\sigma^4}\mathsf{E} X_k \xi^T \xi X_k^T = \frac{1}{\sigma^4}\mathsf{E}(\xi, X_k)^2 = \frac{|X_k|^2 \mathsf{E}|\xi|^2}{\sigma^4} = \frac{|X_k|^2}{\sigma^2},$$

$$I_{r+1,r+1}(\theta) = \frac{1}{4\sigma^4}\mathsf{E}\left[\sum_{i=1}^{n}\left(\frac{\xi_i^2}{\sigma^2} - 1\right)\right]^2 = \frac{n}{2\sigma^4}, \quad I_{ij}(\theta) = 0$$

pour $i \neq j$. De sorte que

$$I^{-1}(\theta) = \left\| \begin{array}{ccc|c} \frac{\sigma^2}{|X_1|^2} & & 0 & 0 \\ & \ddots & \vdots & \\ 0 & & \frac{\sigma^2}{|X_r|^2} & 0 \\ \hline 0 & \dots & 0 & \frac{2\sigma^4}{n} \end{array} \right\|.$$

Donc, dans l'inégalité de Rao-Cramer

$$\mathsf{E}_\theta(\theta^* - \theta)^T(\theta^* - \theta) \geqslant I^{-1}(\theta) \tag{16}$$

l'égalité est réalisée pour les r premières composantes de θ^*. Elle ne l'est pas pour la $(r + 1)$-ième (bien que les deux parties de (16) se conduisent asymptotiquement de la même manière), puisque la condition nécessaire et suffisante du théorème 2.16.1 A est mise en défaut.

REMARQUE 3. La condition de normalité des ξ_i devient de peu d'importance pour les propositions (10), (11) et (12) si n est grand (dans (11) il est préférable de procéder à une normalisation et d'affirmer que la variable aléatoire suit une loi approximativement normale).

REMARQUE 4. Le terme « régression » concerne la distribution conjointe de deux variables aléatoires ξ et η et désigne la courbe

$$g(x) = \mathsf{E}(\eta | \xi = x)$$

qui s'appelle également courbe d'estimation ou de régression de η en ξ. Si par exemple $(\xi, \eta) \in \Phi_{\gamma, \sigma^2}$, $\gamma = (\gamma_1, \gamma_2)$, $\sigma^2 = \|\sigma_{ij}\|$, $i, j = 1, 2$, alors $g(x) = \gamma_2 + \frac{\sigma_{12}}{\sigma_{22}}(x - \gamma_1)$ comme on l'a vu dans les chapitres précédents. Ceci est une régression linéaire élémentaire.

REMARQUE 5. L'hypothèse que les ξ_i suivent la même loi Φ_{0,σ^2}, σ^2 étant connue, peut être affaiblie. On peut admettre que $\xi_i \in \Phi_{0,\sigma_i^2}$ si les σ_i sont différentes et connues. Dans ce cas, en désignant la matrice diagonale $\begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \end{pmatrix}$ par σ et en introduisant les nouvelles variables $\xi' = \xi\sigma^{-1}$, $X' = X\sigma^{-1}$, $Y' = Y\sigma^{-1}$ (de sorte que $\xi_i' = \xi_i/\sigma_i$, $x_i' = x_i/\sigma_i$, $y_i' = y_i/\sigma_i$), on est conduit au problème de régression

$$Y' = \alpha X' + \xi',$$

dans lequel le vecteur des observations Y' et le régresseur X' sont connus, $\xi' \in \Phi_{0,E_n}$. Il est immédiat de vérifier (nous laissons ceci au soin du lecteur) qu'on a l'analogue suivant du théorème 1.

THÉORÈME 2. *L'estimateur*

$$\alpha^* = Y\sigma^{-2}X^T(D')^{-1}, \quad D' = X\sigma^{-2}X^T,$$

est un estimateur efficace sans biais,

$$(\alpha^* - \alpha)(D')^{1/2} \in \Phi_{0,E_r},$$

$$|Y' - \alpha^* X'|^2 = \sum_{i=1}^{n} \frac{\left(y_i - \sum_{k=1}^{r} \alpha_k^* x_{ik}\right)^2}{\sigma_i^2} \in \mathbf{H}_{n-r}.$$

Considérons encore le théorème 1. Les relations (10), (11) et (12) qui y ont été établies nous permettent de construire les régions de confiance aussi bien pour des coordonnées particulières de θ que pour θ dans son ensemble. Par exemple

$$\mathbf{P}_\theta\left(\frac{(n-r)(\sigma^2)^*}{h_{\epsilon/2}^{(2)}} < \sigma^2 < \frac{(n-r)(\sigma^2)^*}{h_{\epsilon/2}^{(1)}}\right) = 1 - \epsilon, \tag{17}$$

et si X_k sont orthogonaux, alors

$$\mathbf{P}_\theta\left(|\alpha_k - \alpha_k^*| < \frac{t_{\epsilon/2}\sigma^*}{|X_k|}\right) = 1 - \epsilon, \tag{18}$$

où

$$\mathbf{T}_{n-r}(]-t_{\epsilon/2}, t_{\epsilon/2}[) = 1 - \epsilon, \quad \mathbf{H}_{n-r}(]-h_{\epsilon/2}^{(1)}, h_{\epsilon/2}^{(2)}[) = 1 - \epsilon.$$

Supposons que les X_k sont orthogonaux. Désignons par $\bar{\alpha}$ le sous-vecteur de α défini dans le corollaire 1. Le théorème 1 nous recommande de construire la région de confiance pour $\bar{\alpha}$ à l'aide de la relation

$$\frac{|(\bar{\alpha} - \bar{\alpha}^*)\bar{X}|^2}{(n-r)(\sigma^2)^*} < f_\epsilon. \tag{19}$$

La valeur f_ϵ correspondant au niveau donné $1 - \epsilon$ se détermine (comme dans le chapitre 3) à l'aide de la distribution de Fisher $\mathbf{F}_{l,n-r}$ à $(l, n-r)$ degrés de liberté.

Si σ^2 est connue, l'intervalle de confiance sera défini par la relation

$$|(\bar{\alpha} - \bar{\alpha}^*)\bar{X}|^2 < \sigma^2 h_\epsilon, \tag{20}$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution $\mathbf{H}_l$.

Dans les problèmes de régression on peut avoir aussi à estimer la valeur de la surface de régression $y = \alpha z^T$ en un nouveau point $z = (z_1, \ldots, z_r) \in$

$\in R^r$ donné *a priori*. Posons $y^* = \alpha^* z^T$. On trouve comme précédemment

$$y^* - y = (\alpha^* - \alpha)z^T = \xi X^T D^{-1} z^T \in \Phi_{0,d^2},$$

$$d^2 = \sigma^2 z D^{-1} z^T, \quad \frac{y^* - y}{d\sigma^*} \in \mathbf{T}_{n-r}.$$

Ceci nous permet de construire les intervalles de confiance pour y.

Signalons que la recherche de la région de confiance pour la surface de régression « dans l'ensemble » est un problème plus compliqué (comparer avec [73]). L'ensemble des surfaces composant la région de confiance sera défini par la région de confiance de θ, région qui est construite par exemple à l'aide de (10), (11) (cf. § 3.8). Pour plus de détails voir [73].

3. Test d'hypothèses concernant la régression linéaire. Nous aborderons deux types de problèmes.

1) Supposons que l'on sache que la représentation (1), (2) a lieu. On demande de tester l'hypothèse que θ est égal à une valeur donnée θ' ou que l'ensemble de coordonnées $\theta_{k_1}, \ldots, \theta_{k_l}$ est égal à l'ensemble $\theta'_{k_1}, \ldots, \theta'_{k_l}$, les autres coordonnées étant inconnues.

Il est commode de construire les tests de telles hypothèses à l'aide des régions de confiance (17) à (20) (cf. § 3.8). Supposons par exemple qu'on demande de tester l'hypothèse H_1 que Y est indépendant de X pour une régression linéaire élémentaire, c'est-à-dire l'hypothèse $H_1 = \{\alpha_2 = 0\}$. De (18) (ou de (14)) on déduit un test de niveau $1 - \epsilon$ infirmant l'hypothèse H_1 si

$$|\alpha_2^*| \geqslant t_{\epsilon/2}\sigma^*/|X_2|. \tag{21}$$

Dans le cas général d'une régression (1) avec des X_k orthogonaux, l'hypothèse que Y est indépendant de X sera de la forme $H_1 = \{\bar{\alpha} = 0\}$, où $\bar{\alpha} = (\alpha_2, \ldots, \alpha_r)$, $x_{i1} \equiv 1$ et pour l'éprouver on peut se servir du test

$$\frac{|\bar{\alpha}^* \bar{X}|^2}{(n-r)(\sigma^2)^*} \geqslant f_\epsilon, \tag{22}$$

où $\bar{X}$ et f_ϵ sont définis dans (19) pour $l = r - 1$.

On peut appliquer aussi les approches du § 3.15 dans lequel on a testé l'appartenance de la loi de l'échantillon à une sous-famille paramétrique. On est alors conduit à un test du rapport de vraisemblance qui, dans un certain sens, sera proche de (22). Si σ^2 est connue, le test du rapport de vraisemblance de $H_1 = \{\bar{\alpha} = 0\}$ sera de la forme

$$\sigma^{-2}|\bar{\alpha}^* \bar{X}|^2 > h_\epsilon,$$

où h_ϵ est le quantile d'ordre $1 - \epsilon$ de la loi $\mathbf{H}_{r-1}$. Ce test sera minimax (cf. §§ 3.9, 3.10) pour les alternatives séparées en conséquence.

2) Test de l'hypothèse de la présence de la régression (1), (2) dans l'échantillon (X, Y). On sous-entend par là l'hypothèse que la représentation (1), (2) a lieu pour des α et σ quelconques, c'est-à-dire que $\sigma^{-1}(Y - \alpha X) \in \Phi_{0,E_n}$ pour des α et σ quelconques. On reconnaît ici un problème d'appartenance de la loi de Y à une famille paramétrique. Mais, comme déjà signalé, les observations de Y ne suivent pas la même loi. Pour ramener le problème à des observations équidistribuées (cf. § 3.17), on se servira de la proposition suivante qui complète le théorème 1. Admettons que les X_k sont orthogonaux.

THÉORÈME 3. *Soit C une matrice orthogonale d'ordre n dont les r premières colonnes sont les colonnes de la matrice $X^T D^{-1/2}$. Alors les coordonnées du vecteur $\delta = (Y - \alpha^* X)C$ sont indépendantes et telles que $\delta_1 = \ldots = \delta_r = 0$, $\delta_i \in \Phi_{0,\sigma^2}$, $i = r + 1, \ldots, n$.*

Le problème se ramène donc au test de l'hypothèse que la loi de l'échantillon $\delta_{r+1}, \ldots, \delta_n$ de taille $n - r$ appartient à la famille Φ_{0,σ^2} (par abus de langage r observations ont été utilisées pour estimer α). Ce problème a été étudié au § 3.17. Pour déterminer les valeurs δ_i il faut calculer successivement les valeurs α^* et $Y - \alpha^* X$ au vu des échantillons X et Y et appliquer à $Y - \alpha^* X$ toute transformation C jouissant des propriétés signalées dans le théorème 3.

Si σ est connue, on est conduit à un problème de test de l'hypothèse simple de distribution suivant la loi Φ_{0,σ^2}. Cependant pour tester l'hypothèse qui nous intéresse dans ce cas, on peut solliciter le théorème 1 qui dit que

$$(n - r)(\sigma^2)^* / \sigma^2 \in \mathbf{H}_{n-r}.$$

DÉMONSTRATION du théorème 3. Si $Z \perp \mathscr{L}[X]$, les r premières coordonnées du vecteur ZC forment le vecteur $ZX^T D^{-1/2} = 0$. Comme $(Y - \alpha^* X) \perp \mathscr{L}[X]$ et $\delta = (Y - \alpha^* X)C$, on en déduit que $\delta_1 = \ldots = \delta_r = 0$. Par ailleurs

$$\delta = (Y - \alpha X)C - (\alpha^* - \alpha)XC = \eta - \bar{\eta} D^{-1/2} XC,$$

où $\eta = \xi C$, $\bar{\eta} = (\eta_1, \ldots, \eta_r) = (\alpha^* - \alpha)D^{1/2} = \xi X^T D^{-1/2}$ et par suite δ est l'image de η par une transformation linéaire,

$$|\delta|^2 = |Y - \alpha^* X|^2 = |\xi|^2 - |(\alpha - \alpha^*)X|^2 =$$

$$= \sum_{i=1}^{n} \eta_i^2 - |\bar{\eta}|^2 = \sum_{i=r+1}^{n} \eta_i^2,$$

de sorte que $\sum_{i=r+1}^{n} \delta_i^2 = \sum_{i=r+1}^{n} \eta_i^2$. Ceci n'est possible que si $(\delta_{r+1}, \ldots, \delta_n)$ est l'image du vecteur $(\eta_{r+1}, \ldots, \eta_n)$ par une rotation ou, ce qui est équivalent, par une transformation orthogonale. Ce qui prouve le théorème, puisque $\sigma^{-1}\eta \in \Phi_{0,E_n}$.

EXEMPLE 1. Dans cet exemple on se propose de décrire l'aspect mathématique d'une expérience physique qui a permis de découvrir la désintégration d'un méson φ en deux mésons π (cf. [74]). Le résultat obtenu revêt un caractère statistique et utilise en fait un modèle de régression.

On étudie l'interaction d'électrons (e^-) et de positrons (e^+) se déplaçant à la rencontre les uns des autres. Si l'énergie totale $2E$ de ces particules se trouve au voisinage du point $2E_0 = 1019{,}6$ MeV (fig. 8), leur collision

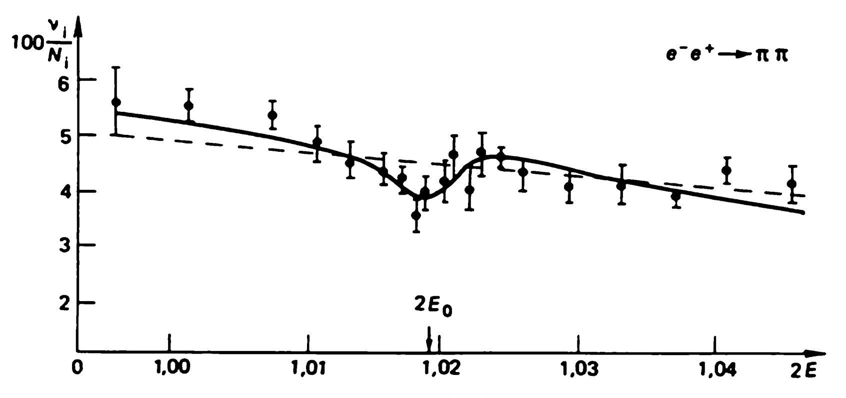


Fig. 8. Courbes représentatives des estimations des lignes de régression sous les hypothèses H_1 et H_2

engendre (entre autres) des particules de deux types : des mésons φ et des couples de mésons π. La probabilité d'apparition d'un couple de mésons π par interaction de e^+ et e^- se décrit à l'aide de E avec une grande précision par une fonction linéaire que nous représentons sous la forme (hypothèse H_1)

$$p_1^{\pi\pi}(E) = a_0 + a_1 x, \quad x = E - E_0, \tag{23}$$

où a_0 et a_1 sont inconnus.

On a avancé l'hypothèse (H_2) que la désintégration des mésons φ engendrés peut également faire apparaître des couples de mésons π. Il est impossible de déceler directement cet effet, car il a été établi que si ce phénomène se produit, il se produit très rarement, pas plus d'une fois toutes les 10^4 désintégrations des mésons φ. Mais grâce à l'interférence de cette source supplémentaire de naissance de mésons π avec la source principale,

la probabilité d'apparition de ces particules sera égale non pas à (23), mais à

$$p_2^{\pi\pi}(E) = [a_0 + a_1 x]\left[1 + \frac{b_0 + b_1 x + b_2 x^2}{x^2 + d^2}\right] \tag{24}$$

(tout comme (23), la relation (24) est une approximation assez exacte d'une formule plus compliquée, approximation qui est basée sur le fait que la marge de variation de $x = E - E_0$ est petite en regard de E_0). Dans cette égalité, les coefficients b_i et a_i sont inconnus, d, connu.

Pour déterminer laquelle des deux relations (23) ou (24) a effectivement lieu, on a réalisé $n = 20$ expériences pour diverses valeurs de l'énergie $E_1, \ldots, E_{20}$.

Tableau 1. **Tableau des données expérimentales**

Numéro de l'observation	E_i, MeV	N_i	ν_i	Numéro de l'observation	E_i, MeV	N_i	ν_i
1	497,75	6 960	384	11	510,40	14 322	716
2	500,65	7 908	435	12	510,92	13 470	568
3	503,65	8 102	432	13	511,39	12 008	569
4	505,40	22 259	1080	14	512,17	23 951	1117
5	506,62	16 938	765	15	513,20	27 796	1185
6	507,66	21 728	951	16	514,62	37 771	1539
7	508,40	14 014	603	17	516,58	25 902	1036
8	508,90	13 793	545	18	518,64	27 857	1057
9	509,40	14 075	615	19	520,61	23 228	989
10	509,90	14 867	691	20	522,88	26 482	1066

Le but de ces expériences (cf. tableau 1 et fig. 8) est de déterminer les nombres N_i, $i = 1, \ldots, 20$, d'interactions de e^+ et e^- et les nombres ν_i de couples de mésons π nés pour une valeur E_i de l'énergie. Les nombres N_i et ν_i sont assez grands (N_i sont de l'ordre de 10^4). Vu que pour N_i fixe, le nombre ν_i des couples de mésons π est distribué suivant la loi de Bernoulli $\mathbf{B}_{p_i}^{N_i}$ ($p_i = p_1^{\pi\pi}(E_i)$ pour l'hypothèse H_1 et $p_i = p_2^{\pi\pi}(E_i)$ pour l'hypothèse H_2), en se servant de l'approximation normale, on est en droit de considérer que

$$y_i \equiv \frac{\nu_i}{N_i} = p_i + \xi_i, \quad \xi_i \in \Phi_{0,\sigma_i^2}$$

(dans le terme ξ_i figurent aussi des perturbations aléatoires (le fond)). En vertu de (23) et de (24), on a deux régressions possibles :

$$p_i = \sum_{k=0}^{1} \alpha_k \psi_k(x_i), \quad \psi_k(x) = x^k, \quad k = 0, 1 \tag{25}$$

(l'hypothèse H_1) et

$$p_i = \sum_{k=0}^{3} \alpha_k \psi_k(x_i), \quad \psi_k(x) = \frac{x^k}{x^2 + d^2}, \quad k = 0, 1, 2, 3 \tag{26}$$

(l'hypothèse H_2).

Les valeurs σ_i^2 varient très peu si les hypothèses sont modifiées. Ces valeurs peuvent être estimées avec une grande précision et l'on peut admettre qu'elles sont connues. Le théorème 2 affirme que la statistique

$$\chi^2 = \left|Y' - \sum_k \alpha_k^* \psi_k\right|^2 = \sum_{i=1}^{n} \left(y_i - \sum_k \alpha_k^* \psi_k(x_i)\right)^2 / \sigma_i^2 \tag{27}$$

suit la distribution $\mathbf{H}_{n-r}$, où r est le nombre de paramètres α_k estimés ($r = 2$ pour l'hypothèse H_1 et $r = 4$ pour l'hypothèse H_2).

Les calculs effectués conformément aux recommandations du théorème 2 nous donnent les valeurs suivantes pour la statistique (27) : $\chi_1^2 = 36{,}8$ dans le premier cas ($r = 2$) et $\chi_2^2 = 19{,}0$ dans le deuxième cas ($r = 4$). Les niveaux réellement atteints (cf. § 3.4) du test $\chi^2 > c$ des hypothèses (de base) H_1 et H_2 sont respectivement égaux à $\mathbf{H}_{18}(]0, 36{,}8[) = 0{,}9944$, $\mathbf{H}_{16}(]0, 19{,}0[) = 0{,}731$.

En d'autres termes, l'hypothèse de l'absence d'une source supplémentaire de génération des couples de mésons π est rejetée par un test basé sur la statistique du χ^2, de niveau 0,99 par exemple. Dans le même temps, l'hypothèse de la présence de cette source s'accorde bien avec les données expérimentales.

Pour être plus exacts, dans ce problème nous aurions dû tester deux hypothèses paramétriques multiples correspondant à (25) et (26) pour les valeurs des probabilités d'apparition de couples de mésons π. Si l'on fait appel à un test du rapport de vraisemblance, on s'assure immédiatement qu'il sera basé sur la différence des statistiques du χ^2 correspondant aux modèles (25) et (26) et donc il fournira des résultats à peu près identiques.

4. Estimation et test d'hypothèses en présence de liaisons linéaires. Considérons comme précédemment une régression linéaire (1), (2) dans l'hypothèse que les coordonnées du vecteur α sont liées par $s < r$ relations linéaires

$$\sum_{k=1}^{r} \alpha_k a_{kl} = c_l, \quad l = 1, \ldots, s.$$

Ces relations peuvent être transcrites sous la forme matricielle

$$\alpha A = c, \tag{28}$$

où A est une $(r \times s)$-matrice que l'on supposera être de rang s. Nous aurions pu dans ce cas exprimer s variables (disons $\alpha_{r-s+1}, \ldots, \alpha_r$) en fonction des autres (c'est-à-dire de $\alpha_1, \ldots, \alpha_{r-s}$), porter les valeurs acquises dans (1), (2) et obtenir de nouveau un problème standard de régression linéaire (mais avec un autre régresseur).

Pour la suite il nous sera plus commode d'aborder la résolution de ce problème sous un autre angle. Adressons-nous à la démonstration du théorème 1. Le sous-espace $\mathscr{A}$ des valeurs α, défini par (28) induit dans $\mathscr{L}[X]$ un sous-espace de dimension s des valeurs αX qui sera désigné par $\mathscr{L}_A[X]$. Il est évident que $\alpha \in \mathscr{A}$ peut être estimé par les méthodes du théorème 1. L'estimateur cherché $\alpha_A^* \in \mathscr{A}$ sera défini comme dans le théorème 1 à l'aide du projeté $\alpha_A^* X$ de Y sur $\mathscr{L}_A[X]$. Donc, conjointement à $(Y - \alpha^* X) \perp \perp \mathscr{L}[X]$ nous obtiendrons la relation $(Y - \alpha_A^* X) \perp \mathscr{L}_A[X]$ qui définit α_A^* de façon unique. Pour déterminer α_A^* il est plus commode d'utiliser une approche analytique : appliquer la méthode des multiplicateurs indéterminés de Lagrange pour calculer $\min_\alpha |Y - \alpha X|^2$ sous la condition $\alpha A = c$.

A cet effet nous devons résoudre les équations

$$\alpha A = c, \quad \frac{\partial}{\partial \alpha}\left[|Y - \alpha X|^2 + \lambda(\alpha A - c)^T\right] = 0 \tag{29}$$

(nous utilisons les multiplicateurs $\lambda_1, \ldots, \lambda_s$ qui forment le vecteur λ et qui correspondent aux conditions (28)). Vu que $|Y - \alpha X|^2 = (Y - \alpha X)(Y - - \alpha X)^T$, la deuxième équation (29) devient

$$-2YX^T - 2\alpha XX^T + \lambda A^T = 0.$$

D'où

$$\alpha_A^* = YX^T D^{-1} - \frac{1}{2}\lambda A^T D^{-1} = \alpha^* - \frac{1}{2}\lambda A^T D^{-1}.$$

D'après (29) on a $c = \alpha_A^* A = \alpha^* A - \frac{1}{2}\lambda A^T D^{-1} A$. La matrice D étant définie positive et A étant de rang s, la matrice $B = D^{-1/2}A$ sera aussi de rang s et la matrice $B^T B = A^T D^{-1} A$ sera aussi définie positive (cf. n° 1). Donc

$$-\frac{1}{2}\lambda = (c - \alpha^* A)\, D_A,$$

$$\alpha_A^* = \alpha^* + (c - \alpha^* A)\, D_A\, A^T D^{-1}, \tag{30}$$

où pour simplifier nous avons posé $D_A = [A^T D^{-1} A]^{-1}$.

Le lecteur pourra s'assurer que α_A^* est un estimateur du maximum de vraisemblance du paramètre α sous la condition $\alpha A = c$. On pourrait obtenir le même résultat (30) par des considérations géométriques en utilisant la

relation $\alpha_A^* \in \mathscr{L}_A[X]$ et l'orthogonalité

$$\begin{aligned} &(Y - \alpha_A^* X) \perp \mathscr{L}_A[X], \\ (\alpha_A^* - \alpha^*)X = [(Y - \alpha^* X) &- (Y - \alpha_A^* X)] \perp \mathscr{L}_A[X]. \end{aligned} \tag{31}$$

Considérons maintenant le problème de *test d'hypothèses linéaires*. L'hypothèse H_1 concernant le paramètre α sera appelée *linéaire* si elle est de la forme $H_1 = \{\alpha A = c\}$, où les matrices A et c sont définies plus haut.

Signalons d'emblée qu'en introduisant le nouveau paramètre $\beta = \alpha A_c$, où A_c est une matrice non dégénérée quelconque dont les s premières colonnes sont confondues avec celles de A, on ramène le problème à la régression

$$Y = \beta X' + \xi, \qquad X' = A_c^{-1} X, \tag{32}$$

et au test de l'hypothèse $\{\bar{\beta} = c\}, \bar{\beta} = (\beta_1, \ldots, \beta_s)$ (cf. n° 2).

Il est naturel de partir aussi des considérations suivantes. Plus l'écart entre αA et c est grand, plus αX est éloigné de $\mathscr{L}_A[X]$ et plus les points αX et $\alpha^* X$ seront distants de $\alpha_A^* X \in \mathscr{L}_A[X]$. Il est donc naturel de poser à la base du test de l'hypothèse H_1 la distance de $\alpha_A^* X$ à $\alpha^* X$. Si l'hypothèse H_1 est vraie, on a grâce à (31)

$$|(\alpha_A^* - \alpha^*)X|^2 = |Y - \alpha_A^* X|^2 - |Y - \alpha^* X|^2. \tag{33}$$

En vertu de (30) (et en remplaçant c par αA) $\alpha_A^* - \alpha^*$ est l'image de $\alpha - \alpha^*$ par une transformation linéaire. Donc, $(\alpha_A^* - \alpha^*)X$ est indépendant de $Y - \alpha^* X$ (cf. théorème 1).

Par ailleurs, d'après (30)

$$\begin{aligned} |(\alpha_A^* - \alpha^*)X|^2 &= (\alpha_A^* - \alpha^*)XX^T(\alpha_A^* - \alpha^*)^T = \\ &= (c - \alpha^* A)D_A(c - \alpha^* A) = (\alpha^* - \alpha)AD_A A^T(\alpha^* - \alpha)^T. \end{aligned} \tag{34}$$

Comme

$$(\alpha^* - \alpha)A = \xi X^T D^{-1} A \in \Phi_{0,\sigma^2 A^T D_A^{-1}} = \Phi_{0,\sigma^2 D_A^{-1}},$$

il vient d'après (34) et le § 2.2 (n° 4)

$$\frac{1}{\sigma^2} |(\alpha_A^* - \alpha^*)X|^2 \in \mathbf{H}_s. \tag{35}$$

De ce qui précède et du théorème 1 il résulte

$$\frac{|(\alpha_A^* - \alpha^*)X|^2}{|Y - \alpha^* X|^2} = \frac{|Y - \alpha_A^* X|^2}{|Y - \alpha^* X|^2} - 1 \in \mathbf{F}_{s,n-r}. \tag{36}$$

Les relations (35) et (36) nous permettent de construire des tests (basés sur la distance de $\alpha^* X$ à $\alpha_A^* X$) de l'hypothèse H_1 respectivement dans les cas où σ^2 est connue et inconnue (cf. chap. 3).

Il est important de signaler que l'hypothèse H_1 est une hypothèse que la loi de α appartient à une sous-famille paramétrique (en présence d'un paramètre fantôme σ^2 si σ^2 est inconnu) et les statistiques (35) et (36) ne sont autres que des statistiques du rapport de vraisemblance (cf. §§ 3.10, 3.15). En effet, supposons par exemple que σ^2 est inconnu. Alors (cf. (5), (8))

$$\sup_{\alpha,\sigma} f_\theta(Y) = \sup_{\alpha,\sigma} (\sqrt{2\pi}\sigma)^{-n} \exp\left\{-\frac{|Y-\alpha X|^2}{2\sigma^2}\right\} =$$

$$= (\sqrt{2\pi}\,\hat{\sigma}^*)^{-n} \exp\left\{-\frac{|Y-\alpha^* X|^2}{2(\sigma^2)^*}\right\} = \left(\sqrt{2\pi}\,\frac{|Y-\alpha^* X|}{n}\right)^{-n} e^{-n/2}.$$

La valeur $\sup\limits_{\alpha\in\mathscr{A},\sigma} f_\theta(Y)$ se calcule exactement de la même façon. On remarquera simplement que si $\alpha \in \mathscr{A}$, l'estimateur du maximum de vraisemblance de α est α_A^* et celui de σ^2 est, comme dans (8), $\frac{1}{n}\,|Y - \alpha_A^* X|^2$. Donc

$$\sup_{\alpha\in\mathscr{A},\sigma} f_\theta(Y) = \left(\sqrt{2\pi}\,\frac{|Y-\alpha_A^* X|}{n}\right)^{-n} e^{-n/2},$$

$$\frac{\sup\limits_{\alpha\in\mathscr{A},\sigma} f_\theta(Y)}{\sup\limits_{\alpha,\sigma} f_\theta(Y)} = \frac{|Y-\alpha^* X|^n}{|Y-\alpha_A^* X|^n}$$

par conséquent la statistique du test du rapport de vraisemblance est équivalente à (36).

Si σ^2 est connu on peut poser la relation (35) à la base du test de l'hypothèse H_1. Le lecteur pourra s'assurer comme dans ce qui précède que c'est aussi un test du rapport de vraisemblance. Ce test étant invariant par un changement du paramètre (cf. § 3.10), la remarque et les propositions des §§ 3.9, 3.10 nous permettent d'affirmer que le test du rapport de vraisemblance

$$|(\alpha_A^* - \alpha^*)X|^2 > \sigma^2 h_\epsilon,$$

où h_ϵ est le quantile d'ordre $1-\epsilon$ de la distribution $\mathbf{H}_s$, sera un test minimax de niveau $1-\epsilon$ de H_1 contre les alternatives séparées de façon convenable.

Grâce à ce qui précède et aux résultats des chapitres 2 et 3 (voir en particulier § 3.15) on peut considérer que le test (36) et l'estimateur (30) seront aussi optimaux. Nous ne nous attarderons pas sur cette question. Un exposé assez complet des problèmes de régression est accessible dans [73].

§ 4. Analyse de variance

Les problèmes d'analyse de variance développés plus bas sont par essence des problèmes de régression. Ces derniers étudient la dépendance des observations par rapport à un facteur numérique x susceptible de prendre des valeurs quelconques $x_1, \ldots, x_n$ données *a priori* et représentant chacune une observation. Dans les problèmes d'analyse de variance on étudie généralement l'action des seuls facteurs discrets (un, deux ou plusieurs) qui ne peuvent prendre qu'un nombre fini de valeurs. Pour chacune de ces valeurs nous disposons d'un ensemble d'observations (un échantillon). L'analyse de variance regroupe des méthodes statistiques basées sur l'analyse des erreurs quadratiques moyennes et destinées à tester les diverses hypothèses et à estimer les paramètres liés à l'action des facteurs. Les principes de l'analyse de variance ont été posés par Fisher.

1. Problèmes d'analyse de variance traités comme des problèmes de régression. Cas d'un seul facteur. Soient donnés r échantillons indépendants

$$Y_1 = (y_{11}, \ldots, y_{1n_1}), \ldots, Y_r = (y_{r1}, \ldots, y_{rn_r})$$

de tailles $n_1, \ldots, n_r$ prélevés dans des populations normales : $Y_k \in \Phi_{\alpha_k, \sigma^2}$. Supposons que les observations Y_k, $k = 1, \ldots, r$, ont été réalisées pour des valeurs différentes d'un facteur auquel on s'intéresse et qui influe sur les valeurs de la moyenne α_k. La variance σ^2 qui, en principe, est inconnue est supposée être la même pour tous les échantillons. Les problèmes d'analyse de variance comprennent le test d'hypothèses concernant les valeurs $\alpha_1, \ldots$ $\ldots, \alpha_r$ et en particulier de l'hypothèse d'homogénéité $\{\alpha_1 = \ldots = \alpha_r = \alpha\}$ (ce problème a déjà été envisagé au § 1) ainsi que l'estimation des paramètres α_k et de leur variation.

L'analyse de variance possède au même titre que la régression un très vaste champ d'applications notamment en sociologie, agriculture, biologie, médecine. Un problème assez typique relevant, par exemple, de la médecine est la détermination de la dépendance entre le taux de cholestérine contenue dans le sang d'un individu et la profession de ce dernier.

Les problèmes d'analyse de variance formulés ci-dessus sont des cas particuliers de problèmes de régression linéaire. En effet, les observations y_{ki} peuvent être mises sous la forme

$$y_{ki} = \alpha_k + \xi_{ki}, \quad \xi_{ki} \in \Phi_{0,\sigma^2}, \quad k = 1, \ldots, r, \quad i = 1, \ldots, n_k. \tag{1}$$

Formons le vecteur

$$Y = (y_{11}, \ldots, y_{1n_1}; y_{21}, \ldots, y_{2n_2}; \ldots; y_{r1}, \ldots, y_{rn_r})$$

et le vecteur ξ d'après la même règle. Les relations (1) peuvent alors être écrites sous la forme matricielle $Y = \alpha X + \xi$, où X est une $(r \times n)$-matrice,

$n = n_1 + \ldots + n_r$, de la forme

$$X = \begin{bmatrix} 1 & 1 \ldots 1 & 0 & 0 \ldots 0 & & 0 & 0 \ldots 0 \\ 0 & 0 \ldots 0 & 1 & 1 \ldots 1 & & 0 & 0 \ldots 0 \\ 0 & 0 \ldots 0 & 0 & 0 \ldots 0 & \ldots & 0 & 0 \ldots 0 \\ \ldots & \ldots & \ldots & \ldots & & \ldots & \ldots \\ 0 & 0 \ldots 0 & 0 & 0 \ldots 0 & & 1 & 1 \ldots 1 \end{bmatrix}.$$

Il est évident que les lignes de cette matrice (les vecteurs X_j) sont orthogonales. L'hypothèse $H_1 = \{\alpha_1 = \alpha_2 = \ldots = \alpha_r\}$ peut être représentée sous la forme

$$\alpha A = 0, \tag{2}$$

où A est une matrice de dimension $r \times (r - 1)$

$$A = \begin{bmatrix} 1 & 0 \ldots & 0 \\ 0 & 1 \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & 0 \ldots & 1 \\ -1 & -1 \ldots & -1 \end{bmatrix}.$$

Il est évident que A est de rang $r - 1$.

Nous voyons que le test de l'hypothèse de base H_1 de l'analyse de variance n'est autre qu'un problème de test d'une hypothèse linéaire pour la régression.

Voyons de quelles formes sont les estimateurs efficaces de α et σ^2 trouvés dans le théorème 3.1. Ici $|X_k|^2 = n_k$, la matrice $D = XX^T$ d'ordre r est de la forme

$$D = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n_r \end{pmatrix},$$

$$\alpha_k^* = \frac{(Y, X_k)}{(X_k, X_k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ki} \equiv \bar{y}_{k\cdot}, \tag{3}$$

$$(n - r)(\sigma^2)^* = |Y - \alpha^* X|^2 = \sum_{k=1}^{r} \sum_{i=1}^{n_k} (y_{ki} - \bar{y}_{k\cdot})^2 \equiv Q_2(Y).$$

En outre $\alpha_1^*, \ldots, \alpha_r^*, (\sigma^2)^*$ sont indépendantes. Les intervalles de confiance pour les paramètres α et σ^2 et pour les fonctions de α et σ^2 se construisent comme dans le § 3.

Pour tester l'hypothèse linéaire (2), il nous faut calculer aussi l'estimateur du maximum de vraisemblance α_A^* sous la condition (2) (cf. n° 4 du paragraphe précédent). Le moyen le plus simple est d'utiliser l'approche développée au début du n° 4 § 3 et d'exprimer $\alpha_1, \ldots, \alpha_r$ en fonction des variables indépendantes. Ici nous n'avons qu'une seule variable indépendante ; supposons que c'est $\alpha_r = \mu$ et $\alpha_A^* = (\mu^*, \ldots, \mu^*)$, où μ^* minimise

$$|Y - (\mu, \ldots, \mu)X|^2 = \sum_{k=1}^{r} \sum_{i=1}^{n_k} (y_{ki} - \mu)^2.$$

Il est évident que

$$\mu^* = \bar{y} \equiv \frac{1}{n} \sum_{k=1}^{r} \sum_{i=1}^{n_k} y_{ki},$$

$$|Y - \alpha_A^* X|^2 = \sum_{k=1}^{r} \sum_{i=1}^{n_k} (y_{ki} - \bar{y})^2 \equiv Q(Y) =$$

$$= \sum_{k=1}^{r} \sum_{i=1}^{n_k} (y_{ki} - \bar{y} + \bar{y}_{k\cdot} - \bar{y}_{k\cdot})^2 =$$

$$= \sum_{k=1}^{r} \sum_{i=1}^{n_k} (y_{ki} - \bar{y}_{k\cdot})^2 + \sum_{k=1}^{r} n_k(\bar{y}_{k\cdot} - \bar{y})^2$$

(la somme des produits mixtes est nulle, puisque $\sum_{i=1}^{n_k} (y_{ki} - \bar{y}_{k\cdot}) = 0$). Si l'hypothèse H_1 est vraie, en vertu de (3.33), (3) et de la dernière égalité on a

$$|(\alpha_A^* - \alpha^*)X|^2 = Q(Y) - Q_2(Y) = \sum_{k=1}^{r} n_k(\bar{y}_{k\cdot} - \bar{y})^2 \equiv Q_1(Y).$$

Si l'hypothèse H_1 est vraie, il vient d'après (3.36) que $Q_1(Y)/Q_2(Y) \in$ $\in \mathbf{F}_{r-1,n-r}$, ce qui permet de construire le test $Q_1(Y)/Q_2(Y) > f_\epsilon$ (f_ϵ est le quantile d'ordre $1 - \epsilon$ de $\mathbf{F}_{r-1,n-r}$) de l'hypothèse H_1 qui sera un test du rapport de vraisemblance. Si σ^2 est connue, le test du rapport de vraisemblance sera de la forme

$$Q_1(Y) > \sigma^2 h_\epsilon$$

(h_ϵ est le quantile d'ordre $1 - \epsilon$ de $\mathbf{H}_{r-1}$) et sera un test minimax pour des alternatives convenablement séparées (cf. § 3.9).

2. Influence de deux facteurs. Approche élémentaire. Dans les problèmes de ce numéro on étudie l'influence de deux types de facteurs sur les résultats de l'expérience. En agriculture, par exemple, ce peut être l'étude de l'influence, sur la récolte, de la composition du sol (le facteur A prend r valeurs) et de la méthode de traitement (le facteur B prend s valeurs).

Les observations peuvent être présentées sous la forme

$$\begin{gathered} y_{kli} = \alpha_{kl} + \xi_{kli}, \quad \xi_{kli} \in \Phi_{0,\sigma^2}, \\ k = 1, \ldots, r, l = 1, \ldots, s, i = 1, \ldots, n_{kl}, \end{gathered} \tag{4}$$

et le modèle envisagé ne se distingue en rien du modèle (1) du n° 1. Il est donc justiciable de tous les résultats du § 3, cependant leur application directe soulève de grosses difficultés. La présence des triples indices est en soi une source de difficultés. Pour simplifier un peu le problème, on pose $n_{kl} \equiv 1$; ceci nous débarassera d'un indice (l'indice i dans (4)). Par ailleurs, on proposera une approche élémentaire légèrement différente qui nous permettra d'établir sans recourir aux théorèmes du § 3 des assertions indispensables au test des hypothèses de base.

On étudie donc un échantillon $Y_{kl} = y_{kl}$ de taille un, de sorte que l'ensemble des données empiriques Y se représente par une matrice de $r \times s$ nombres y_{kl} qui sont les résultats de l'expérience réalisée sous l'influence de la k-ième valeur du facteur A et de la l-ième valeur du facteur B. Cette matrice peut être traitée comme une matrice composée de r échantillons (lignes) de taille s correspondant aux diverses valeurs du facteur A ou composée de s échantillons (colonnes) de taille r correspondant aux diverses valeurs du facteur B. Dans la suite le groupement des observations sera réalisé en conséquence. Posons

$$\bar{y}_{k\cdot} = \frac{1}{s} \sum_{l=1}^{s} y_{kl}, \quad \bar{y}_{\cdot l} = \frac{1}{r} \sum_{k=1}^{r} y_{kl}, \quad \bar{y} = \frac{1}{rs} \sum_{k,l} y_{kl}.$$

On a l'identité

$$Q(Y) \equiv \sum_{k,l} (y_{kl} - \bar{y})^2 = Q_1(Y) + Q_2(Y) + Q_3(Y), \tag{5}$$

où

$$Q_1(Y) = s \sum_{k} (\bar{y}_{k\cdot} - \bar{y})^2, \quad Q_2(Y) = r \sum_{l} (\bar{y}_{\cdot l} - \bar{y})^2,$$

$$Q_3(Y) = \sum_{k,l} (y_{kl} - \bar{y}_{k\cdot} - \bar{y}_{\cdot l} + \bar{y})^2.$$

Nous admettrons que l'influence des facteurs est additive, c'est-à-dire qu'il existe des a_k et b_l tels que

$$\alpha_{kl} = a_k + b_l, \quad k = 1, \dots, r, \ l = 1, \dots, s. \tag{6}$$

Il est évident que Q_1 définit les variations de a_k (c'est-à-dire est lié au facteur A), Q_2, les variations de b_l (est lié au facteur B) et Q_3 une somme engendrée entièrement par le hasard. Il est évident par ailleurs que

$$Q_i(Y + a) = Q_i(Y), \quad i = 1, 2, 3. \tag{7}$$

THÉORÈME 1. 1)

$$Q_3(Y)/\sigma^2 \in \mathbf{H}_{(r-1)(s-1)}. \tag{8}$$

2) *Si l'hypothèse* $H_A = \{a_1 = \dots = a_r = a\}$ *est vraie, alors* $Q_1(Y)$ *ne dépend pas de* $Q_2(Y)$ *et de* $Q_3(Y)$, $Q_1(Y)/\sigma^2 \in \mathbf{H}_{r-1}$. *On a une proposition analogue relativement à* Q_2 *et à l'hypothèse* $H_B = \{b_1 = \dots = b_s = b\}$.

3) *Si l'hypothèse* $H_1 = \{\alpha_{kl} = \alpha\}$ *est vraie, les trois formes quadratiques* Q_1, Q_2 *et* Q_3 *sont indépendantes.*

DÉMONSTRATION. Posons sans nuire à la généralité $\sigma^2 = 1$. Alors

$$\mathbf{E} y_{kl} y_{ij} = \begin{cases} \alpha_{kl}\alpha_{ij} & \text{si} \quad (i,j) \neq (k,l), \\ \alpha_{kl}^2 + 1 & \text{si} \quad (i,j) = (k,l). \end{cases}$$

D'où

$$\mathbf{E}\Big(\sum_{\mathrm{I}} y_{kl}\Big)\Big(\sum_{\mathrm{II}} y_{kl}\Big) = \Big(\sum_{\mathrm{I}} \alpha_{kl}\Big)\Big(\sum_{\mathrm{II}} \alpha_{kl}\Big) + m,$$

où m est le nombre de termes semblables dans les sommes $\sum_{\mathrm{I}}$ et $\sum_{\mathrm{II}}$. En se servant de cette égalité, on trouve sans peine que

$$\mathbf{E}(\bar{y}_{k\cdot} - \bar{y})(\bar{y}_{\cdot l} - \bar{y}) = (\alpha_{k\cdot} - \bar{\alpha})(\alpha_{\cdot l} - \bar{\alpha}) = (a_k - \bar{a})(b_l - \bar{b}) \tag{9}$$

sous les conventions naturelles relatives aux notations $\alpha_{k\cdot}$, $\alpha_{\cdot l}$, $\bar{\alpha}$, $\bar{a}$, $\bar{b}$. Si l'hypothèse $H_A = \{a_1 = \dots = a_r = a\}$ est vraie, l'espérance mathématique de (9) est nulle. Vu que $\mathbf{E}(\bar{y}_{k\cdot} - \bar{y}) = \alpha_{k\cdot} - \bar{\alpha} = 0$, ceci exprime que l'ensemble des variables aléatoires $\{\bar{y}_{k\cdot} - \bar{y}\}$ est indépendant de $\{\bar{y}_{\cdot l} - \bar{y}\}$.

On établit de façon analogue que pour tous k, l et i

$$\mathbf{E}(y_{kl} - \bar{y}_{k\cdot})(\bar{y}_{i\cdot} - \bar{y}) = 0.$$

Ceci exprime que l'ensemble $\{\bar{y}_{k\cdot} - \bar{y}\}$ est indépendant aussi de $\{y_{kl} - \bar{y}_{k\cdot} - \bar{y}_{\cdot l} + \bar{y}\}$. Ceci exprime à son tour que si H_A est vraie, $Q_1(Y)$ ne

dépend pas de $Q_2(Y)$ et $Q_3(Y)$. Le fait que $Q_1(Y) \in \mathbf{H}_{r-1}$ résulte du lemme de Fisher (§ 2.32).

La situation est la même si est réalisée l'hypothèse H_B. Si l'hypothèse H_1 est vraie (c'est-à-dire si les hypothèses H_A et H_B ont lieu), il est évident que les trois ensembles de variables aléatoires cités plus haut seront indépendants. Ce qui exprime l'indépendance de $Q_1(Y)$, $Q_2(Y)$ et $Q_3(Y)$.

Reste à trouver la distribution de $Q_3(Y)$. Etant donné que cette distribution ne dépend pas de a_k et b_l, on peut admettre que $a_k = b_l = 0$ pour tous les k et l et que par conséquent l'hypothèse H_1 est réalisée. De la définition de $Q(Y)$ il s'ensuit alors que $Q(Y) \in \mathbf{H}_{rs-1}$. Par ailleurs la relation (5), où $Q_1(Y) \in \mathbf{H}_{r-1}$ et $Q_2(Y) \in \mathbf{H}_{s-1}$, est valable. Reste à utiliser l'indépendance de $Q_i(Y)$ et le lemme 3.1. ◄

On pourrait appliquer la même approche pour les problèmes du n° 1.

Le théorème 1 légitime les procédures statistiques suivantes :

1) Estimation des paramètres $a_k - a_i$, $b_l - b_j$, σ^2 (les nombres a_k et b_l de (6) sont définis à un facteur additif constant près) à l'aide des estimateurs $\bar{y}_{k\cdot} - \bar{y}_{i\cdot}$, $\bar{y}_{\cdot l} - \bar{y}_{\cdot j}$, $(\sigma^2)^* = Q_3(Y)/(r-1)(s-1)$. Ces estimateurs seront efficaces, puisque les raisonnements sont les mêmes que ceux produits dans le § 3 et dans le n° 1 de ce paragraphe. Les intervalles de confiance pour σ^2 et $a_k - a_i$ peuvent être construits à l'aide des relations (8)

$$\bar{y}_{k\cdot} - \bar{y}_{i\cdot} - (a_k - a_i) \in \mathbf{\Phi}_{0,2\sigma^2/s},$$

$$\frac{\bar{y}_{k\cdot} - \bar{y}_{i\cdot} - (a_k - a_i)}{\sqrt{\dfrac{2Q_3(Y)}{s(r-1)(s-1)}}} \in \mathbf{T}_{(r-1)(s-1)}$$

(tout se passe de la même manière pour $b_l - b_j$).

2) Test de l'hypothèse H_A à l'aide du critère $Q_1/Q_3 > f_\epsilon$. Ce test sera de niveau $1 - \epsilon$ si f_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution $\mathbf{F}_{r-1,(r-1)(s-1)}$.

Le test de l'hypothèse H_B sera de la même forme, soit $Q_2/Q_3 > f_\epsilon$, où f_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution $\mathbf{F}_{s-1,(r-1)(s-1)}$.

3) Test de l'hypothèse H_1 à l'aide du critère

$$\frac{Q_1 + Q_2}{Q_3} > f_\epsilon$$

de niveau $1 - \epsilon$, où f_ϵ est le quantile d'ordre $1 - \epsilon$ de la distribution $\mathbf{F}_{r+s-2,(r-1)(s-1)}$.

Les problèmes d'analyse de variance font l'objet d'un examen plus détaillé dans [72], [73].

§ 5. Analyse discriminante

Dans ce paragraphe on survolera un cercle de problèmes relevant de l'analyse discriminante *).

Dans le § 3.1 nous avons envisagé le problème suivant de test de r hypothèses simples. Etant donné les distributions $\mathbf{P}_1, \ldots, \mathbf{P}_r$ et un échantillon X de taille n, on demande de déterminer laquelle des hypothèses

$$H_j = \{X \in \mathbf{P}_j\} \tag{1}$$

est la vraie.

Mais dans les problèmes qui se posent en pratique, les distributions $\mathbf{P}_j$ sont souvent inconnues et l'on ne peut se faire une idée sur elles qu'au vu des échantillons.

Soient donc donnés r échantillons $X_i = (\mathrm{x}_{i1}, \ldots, \mathrm{x}_{in_i})$, $i = 1, \ldots, r$, de tailles respectives $n_1, \ldots, n_r$ associés à r distributions différentes *inconnues* $\mathbf{P}_1, \ldots, \mathbf{P}_r$ et soit donné de plus un échantillon X. On demande de résoudre encore le même problème : déterminer laquelle des hypothèses (1) est vraie. En d'autres termes, il faut dire de quel échantillon $X_1, \ldots, X_r$ l'échantillon X est le prolongement.

Pour simplifier on se bornera à l'étude du cas $r = 2$.

1. Cas paramétrique. Supposons tout d'abord que $\mathbf{P}_i$ appartient à une famille paramétrique $\{\mathbf{P}_\theta\}$ vérifiant la condition (A_μ), c'est-à-dire que $X_1 \in \mathbf{P}_{\theta_1}, X_2 \in \mathbf{P}_{\theta_2}, X \in \mathbf{P}_\theta$ pour certains $\theta_1 \neq \theta_2$, et $\theta = \theta_1$ ou $\theta = \theta_2$. La première de ces égalités correspond à l'hypothèse $H_1 = \{X \in \mathbf{P}_{\theta_1}\}$, la deuxième, à l'hypothèse $H_2 = \{X \in \mathbf{P}_{\theta_2}\}$.

Supposons, toujours par souci de simplicité, que les échantillons sont de même taille : $n_1 = n_2 = n$.

Considérons l'échantillon global (X_1, X_2, X) et représentons-le comme un échantillon de taille n formé par les observations $(\mathrm{x}_{1i}, \mathrm{x}_{2i}, \mathrm{x}_i)$ et distribué suivant la loi $\mathbf{P}_{\theta_1} \times \mathbf{P}_{\theta_2} \times \mathbf{P}_\theta$ de densité $f_{\theta_1}(x_1) f_{\theta_2}(x_2) f_\theta(x)$ dépendant du paramètre $\bar\theta = (\theta_1, \theta_2, \theta)$. Il est évident que la fonction de vraisemblance de l'échantillon (X_1, X_2, X) sera égale à

$$f_{\bar\theta}(X_1, X_2, X) = f_{\theta_1}(X_1) f_{\theta_2}(X_2) f_\theta(X).$$

Nous sommes conduits au problème de test de l'hypothèse H_1 que le paramètre $\bar\theta$ est situé sur la « courbe » $\theta = \theta_1$ contre l'hypothèse H_2 que $\bar\theta$ est situé sur une autre « courbe » $\theta = \theta_2$. Ceci est un problème de test de

*) Les problèmes dans lesquels les distributions $\mathbf{P}_i$ de (1) sont connues relèvent aussi de l'analyse discriminante.

l'hypothèse que la loi de l'échantillon appartient à une sous-famille paramétrique (cf. § 3.15) mais dans le cas où l'hypothèse concurrente est que cette loi appartient à une autre sous-famille paramétrique. Ce problème se traite comme dans le § 3.15, mais il sort du cadre de cet ouvrage pour sa complexité sur le plan technique. On se bornera ici au cas d'un paramètre θ scalaire et on décrira succinctement la teneur de ce résultat qui est tout à fait identique à celle du § 3.15 : si le paramètre θ est localisé, c'est-à-dire si les points θ_1 et θ_2 sont situés au voisinage d'un point θ_0, $|\theta_1 - \theta_2| > b/\sqrt{n}$ et si la famille $\{\mathbf{P}_\theta\}$ satisfait au point θ_0 les conditions de régularité (RR), alors le test du rapport de vraisemblance

$$\frac{\sup\limits_{\theta_1,\theta_2} f_{\theta_1}(X_1) f_{\theta_2}(X_2) f_{\theta_2}(X)}{\sup\limits_{\theta_1,\theta_2} f_{\theta_1}(X_1) f_{\theta_2}(X_2) f_{\theta_1}(X)} > c \tag{2}$$

sera asymptotiquement minimax de H_1 contre H_2 pour $n \to \infty$.

La restriction $n_1 = n_2 = n$ n'est pas essentielle. On peut s'en dédouaner comme dans le § 1.

2. Cas général. Dans le cas général où l'on n'a aucune raison de supposer que les X_i sont liés à une famille paramétrique, on peut développer une approche générale basée sur les mêmes considérations que celles qui ont été utilisées pour construire des tests d'homogénéité dans le § 2. Dans ce cas le test de H_1 contre H_2 sera une fonction de trois échantillons, de sorte que $\pi = \pi(X_1, X_2, X)$ est la probabilité d'accepter H_2 au vu de (X_1, X_2, X). Comme précédemment le test non randomisé est défini par sa région critique $\Omega \subset \mathscr{X}^{n_1+n_2+n}$ dans l'espace des échantillons (X_1, X_2, X). Il est naturel d'appeler niveau ou seuil de signification de ce test, le nombre

$$1 - \epsilon = \inf_{\mathbf{P}_1 \in \mathscr{P}, \mathbf{P}_2 \in \mathscr{P}} \mathbf{P}_1 \times \mathbf{P}_2 \times \mathbf{P}_1((X_1, X_2, X) \notin \Omega),$$

où $\mathscr{P}$ est la classe des distributions admissibles. La valeur

$$\beta_\pi(\mathbf{P}_1, \mathbf{P}_2) = \mathbf{P}_1 \times \mathbf{P}_2 \times \mathbf{P}_2((X_1, X_2, X) \in \Omega),$$
$$\mathbf{P}_1 \in \mathscr{P}, \quad \mathbf{P}_2 \in \mathscr{P},$$

est la puissance de ce test au point $(\mathbf{P}_1, \mathbf{P}_2)$.

On dit qu'un test π est convergent (ou consistant) si $\beta_\pi(\mathbf{P}_1, \mathbf{P}_2) \to 1$ lorsque $n_1 \to \infty$, $n_2 \to \infty$, $n \to \infty$ quelles que soient $\mathbf{P}_1 \neq \mathbf{P}_2$, $\mathbf{P}_1 \in \mathscr{P}$, $\mathbf{P}_2 \in \mathscr{P}$.

Pour construire des tests convergents on peut se servir de la distance des distributions empiriques $\mathbf{P}^*_{X_1}$ et $\mathbf{P}^*_{X_2}$ à $\mathbf{P}_1$ et $\mathbf{P}_2$ respectivement. Si $d(\mathbf{P}, \mathbf{Q})$ est une distance entre ces distributions, la distance $d(\mathbf{P}^*_{X_2}, \mathbf{P}^*_X)$ doit être plus petite que $d(\mathbf{P}^*_{X_1}, \mathbf{P}^*_X)$ pour l'hypothèse H_2. Donc, pour critère on peut

prendre l'inégalité

$$d(\mathbf{P}^*_{X_2}, \mathbf{P}^*_X) - d(\mathbf{P}^*_{X_1}, \mathbf{P}^*_X) < c,$$

dont la réalisation exprime que H_2 est vraie. Le calcul de tels tests (le calcul de leurs niveaux et de leurs puissances) est lié généralement à de grosses difficultés (comparer avec les problèmes plus simples du § 2).

Le *groupement des données* nous permet d'utiliser un test asymptotiquement optimal (2) dans le cas général. Supposons qu'on a procédé à un groupement sur les domaines $\Delta_1, \ldots, \Delta_m$ et que $(\nu_{i1}, \ldots, \nu_{im})$ et $(\nu_1, \ldots, \nu_m)$ sont les fréquences d'accès des observations de X_i, $i = 1, 2$, et de X respectivement à ces domaines. Supposons par ailleurs que $\theta_i = (\theta_{i1}, \ldots, \theta_{im})$ sont les probabilités $(\mathbf{P}_i(\Delta_1), \ldots, \mathbf{P}_i(\Delta_m))$ d'accès aux domaines $\Delta_1, \ldots, \Delta_m$ pour les distributions $\mathbf{P}_i$, $i = 1, 2$. Vu que la fonction de vraisemblance $f_{\theta_i}(X_i)$ pour l'échantillon groupé X_i, $i = 1, 2$, est égale à $f_{\theta_i}(X_i) = \prod_{k=1}^{m} \theta_{i_k}^{\nu_{ik}}$, le test (2) sera de la forme

$$\sup_{\theta_2} \sum_{k=1}^{m} (\nu_{2k} + \nu_k) \ln \theta_{2k} + \sup_{\theta_1} \sum_{k=1}^{m} \nu_{1k} \ln \theta_{1k} -$$

$$- \sup_{\theta_1} \sum_{k=1}^{m} (\nu_{1k} + \nu_k) \ln \theta_{1k} - \sup_{\theta_2} \sum_{k=1}^{m} \nu_{2k} \ln \theta_{2k} > \ln c,$$

ou

$$\sum_{k=1}^{m} (\nu_{2k} + \nu_k) \ln \frac{\nu_{2k} + \nu_k}{n_2 + n} + \sum_{k=1}^{m} \nu_{1k} \ln \frac{\nu_{1k}}{n_1} >$$

$$> \ln c + \sum_{k=1}^{m} (\nu_{1k} + \nu_k) \ln \frac{\nu_{1k} + \nu_k}{n_1 + n} + \sum_{k=1}^{m} \nu_{2k} \ln \frac{\nu_{2k}}{n_2}. \quad (3)$$

On peut reproduire les mêmes raisonnements pour $r > 2$.

CHAPITRE 5

LA THÉORIE DES JEUX DANS LES PROBLÈMES DE STATISTIQUE MATHÉMATIQUE

§§ 1, 2 et 3 : notions de jeux ordinaire et statistique. Principales classes de stratégies optimales.
§§ 4, 5 : méthodes de recherche des solutions optimales.
§§ 6, 7 et 8 : construction des décisions asymptotiquement optimales.

§ 1. Remarques préliminaires

Dans les chapitres précédents, nous avons étudié un grand nombre de problèmes de tout genre. Le trait commun de ces problèmes est que le statisticien doit prendre une décision au vu des données expérimentales. En théorie de l'estimation, il doit décider de l'estimation ponctuelle θ^* qu'il faut prendre pour valeur inconnue du paramètre θ, en théorie de tests d'hypothèses statistiques, de la forme des assertions indiquant lesquelles des hypothèses avancées sur la nature de l'événement étudié sont justes, lesquelles sont fausses. Ces décisions, si elles sont erronées, se soldent par des pertes. Une erreur commise dans l'estimation en laboratoire (au vu d'un échantillon) de la composition d'un minerai peut compromettre le régime optimal de coulée et détériorer la qualité du métal. Ceci entraîne de grosses pertes matérielles dont l'ampleur dépend de la gravité de l'erreur. Une fausse décision concernant l'effet d'un remède testé sur un groupe d'expérience de malades peut de toute évidence se solder par des pertes qu'il sera possible d'exprimer par des unités. Nous adopterons cette convention pour les autres problèmes de statistique dans lesquels les dommages ne sont pas explicitement chiffrables.

Ceci nous suggère de distinguer les quatre éléments suivants en statistique mathématique, des éléments qui définissent à proprement parler l'essence de chaque problème concret. Par souci de simplicité nous n'envisagerons dans la suite de l'exposé que des *problèmes portant sur un seul échantillon X de taille n donnée.*

1) Un ensemble Θ dont les éléments θ définissent l'état de l'objet étudié. Si θ est connu, nous n'avons nul besoin de chercher une solution. L'ensemble Θ est appelé aussi ensemble des paramètres bien que θ puisse admettre une plus large interprétation (Θ peut par exemple être très riche et coïncider avec l'ensemble de toutes les distributions sur un espace $\mathscr{X}$).

2) Pour collecter une information sur le paramètre inconnu θ, le statisticien doit effectuer une expérience et procéder à des observations sur une variable aléatoire dont la distribution dépend de θ. En d'autres termes, le statisticien dispose d'un échantillon X distribué suivant une loi $\mathbf{P}_\theta$. Or on sait qu'il est possible d'extraire de l'échantillon X une information sur $\mathbf{P}_\theta$, donc, sur θ. On peut admettre qu'est remplie la condition (A_0) (cf. § 2.6) qui établit une correspondance biunivoque entre θ et $\mathbf{P}_\theta$.

3) Dans les problèmes de statistique, on définit toujours un ensemble $D = \{\delta\}$ de décisions que le statisticien pourra prendre. En théorie de l'estimation, l'ensemble D est généralement confondu avec Θ ; dans les problèmes de test d'hypothèses, l'ensemble D est fini et contient autant d'éléments qu'il y a d'hypothèses à tester. Si θ est connu, la décision $\delta = \varphi(\theta)$ est unique. Si θ est inconnu, il est souhaitable de choisir une décision δ qui soit optimale dans un certain sens. Mais l'optimisation des décisions sous-entend la possibilité de comparer lesdites décisions. Nous admettrons qu'à cet effet est donnée une fonction de perte chiffrant les conséquences d'une prise de décision.

4) La fonction de perte $w(\delta, \theta)$ est définie sur $D \times \Theta$ et indique les pertes qui seront subies si la décision δ est prise lorsque l'objet étudié se trouve dans l'état θ. On conviendra que $w(\delta, \theta) > 0$ pour $\delta \neq \varphi(\theta)$, $w(\varphi(\theta), \theta) = 0$.

Si des quatre éléments précités on retire le n°2 relatif aux données expérimentales, on obtient un *jeu ordinaire à deux joueurs* opposant le statisticien à la nature.

§ 2. Notions fondamentales et théorèmes relatifs au jeu à deux joueurs

1. Jeu à deux joueurs.

DÉFINITION 1. On appelle *jeu à deux joueurs* le triplet (D, Θ, w) composé des ensembles D et Θ et de l'application $w : D \times \Theta \to [0, \infty[$. Les éléments δ de D s'appellent *stratégies du premier joueur*, les éléments $\theta \in \Theta$, *stratégies du second joueur, la fonction w de perte du premier joueur* (ou la fonction de gain du second) définit les pertes $w(\delta, \theta)$ subies par le premier joueur s'il opte pour la stratégie δ et par le second, s'il choisit la stratégie θ.

Le problème fondamental de la théorie des jeux à deux joueurs consiste à choisir une stratégie optimale du premier joueur auquel nous nous identifierons souvent. Il faut à cet effet munir l'ensemble des stratégies d'une certaine relation d'ordre. Cette tâche n'est pas aisée, car les pertes $w(\delta, \theta)$ qui devront être utilisées pour introduire cet ordre dépendent de deux arguments, si bien que pour chaque θ il existera en général une seule stratégie δ qui minimisera $w(\delta, \theta)$.

DÉFINITION 2. On dira qu'une stratégie δ_1 *est meilleure que* δ_2 si

$$w(\delta_1, \theta) \leqslant w(\delta_2, \theta), \quad \forall \theta \in \Theta \tag{1}$$

et s'il existe au moins une valeur $\theta_1 \in \Theta$ telle que $w(\delta_1, \theta_1) < w(\delta_2, \theta_1)$.

Si (1) est seule réalisée, on dira que la stratégie δ_1 *est aussi bonne que* δ_2.

La stratégie δ_0 pour laquelle

$$w(\delta_0, \theta) \leqslant w(\delta, \theta) \quad \textit{pour tous les } \delta \textit{ et } \theta,$$

sera appelée *uniformément optimale* (ou uniformément la meilleure).

La stratégie uniformément la meilleure cause les plus petites pertes quel que soit θ. Mais cette stratégie n'existe généralement pas.

Signalons les trois approches suivantes de l'étude des stratégies optimales du premier joueur :

— recherche des stratégies uniformément optimales dans les sous-classes ;

— recherche des stratégies minimax et bayésiennes ;

— étude de l'ensemble de toutes les stratégies inaméliorables (de la classe complète des stratégies).

2. Stratégies uniformément optimales dans les sous-classes. Dans les problèmes de statistique mathématique, on utilise souvent la démarche suivante (cf. § 5). Pour des raisons de symétrie, de simplicité des calculs, etc., on arrive parfois à restreindre sans perte de généralité la classe des stratégies envisagées. Si cette restriction fait apparaître unc stratégie uniformément optimale, le problème est résolu *ipso facto.* Si l'on applique cette approche, il faut s'assurer nécessairement que la restriction de la classe des stratégies ne nous prive pas de la possibilité d'obtenir une bien meilleure décision. Dans les deux paragraphes suivants, on exhibera des exemples d'application de cette approche à un objet à vrai dire plus compliqué : aux jeux statistiques. Le lecteur en a déjà pris connaissance dans les chapitres 2 et 3 où l'on a étudié les meilleurs estimateurs dans la sous-classe des estimateurs sans biais, et les tests uniformément les plus puissants dans les sous-classes de tests invariants ou sans biais.

3. Stratégies bayésiennes. Elles se présentent dans les cas où le deuxième joueur choisit sa stratégie de façon aléatoire avec une certaine distibution (connu ou inconnue) sur Θ.

Pour pouvoir étudier dans la suite les stratégies « aléatoires » on admettra que Θ et D sont munis des tribus naturelles de sous-ensembles $\mathfrak{F}_\Theta$ et $\mathfrak{F}_D$. On peut alors définir sur $(\Theta, \mathfrak{F}_\Theta)$ et $(D, \mathfrak{F}_D)$ des distributions $\mathbf{Q}$ et π respectivement, de sorte que $(\Theta, \mathfrak{F}_\Theta, \mathbf{Q})$, $(D, \mathfrak{F}_D, \pi)$ seront des espaces probabilisés.

La donnée des distributions π et $\mathbf{Q}$ induit l'espace probabilisé $(D \times \Theta, \mathfrak{F}_{D \times \Theta}, \pi \times \mathbf{Q})$, où $\mathfrak{F}_{D \times \Theta}$ est une tribu engendrée par les produits directs des

ensembles de $\mathfrak{F}_D$ et $\mathfrak{F}_\Theta$. Les tribus $\mathfrak{F}_D$ et $\mathfrak{F}_\Theta$ doivent être choisies de façon à satisfaire les deux conditions suivantes :

a) $\mathfrak{F}_D$ *et* $\mathfrak{F}_\Theta$ *contiennent des singletons* $\{\delta\}$ *et* $\{\theta\}$.

b) *La fonction de perte* $w(\delta, \theta)$ *est mesurable par rapport à* $\mathfrak{F}_{D\times\Theta}$.

DÉFINITION 3. Les distributions π sur $(D, \mathfrak{F}_D)$ et $\mathbf{Q}$ sur $(\Theta, \mathfrak{F}_\Theta)$ seront appelées *stratégies mixtes* ou *randomisées* respectivement du premier et du deuxième joueur.

La distribution $\mathbf{Q}$ sera souvent appelée distribution *a priori*. Ce terme a été expliqué dans les chapitres 2 et 3. Il le sera encore dans le prochain paragraphe.

Désignons par $\tilde{D}$ et $\tilde{\Theta}$ les ensembles de toutes les stratégies mixtes des joueurs 1 et 2 (c'est-à-dire les ensembles de toutes les distributions sur $(D, \mathfrak{F}_D)$ et $(\Theta, \mathfrak{F}_\Theta)$). Vu que $\mathfrak{F}_D$ et $\mathfrak{F}_\Theta$ contiennent des singletons, les ensembles $\tilde{D}$ et $\tilde{\Theta}$ contiendront des distributions concentrées en un point et, par suite, on peut admettre qu'ils renferment des stratégies δ et θ que l'on appellera *stratégies pures* pour les distinguer des autres. On conviendra, sans crainte de confusion, de désigner les distributions de $\tilde{D}$ et $\tilde{\Theta}$ concentrées en un point δ ou θ respectivement par δ et θ.

Les pertes $\tilde{w}(\pi, \mathbf{Q})$ subies en utilisant les stratégies mixtes se définissent par

$$\tilde{w}(\pi, \mathbf{Q}) = \mathbf{E}_{\pi\times\mathbf{Q}} w(\delta, \theta) = \int w(u, t)\pi(du)\mathbf{Q}(dt). \tag{2}$$

Donc, conjointement au jeu primitif nous pouvons envisager un jeu $(\tilde{D}, \tilde{\Theta}, \tilde{w})$ de fonction de perte (2) obtenu par *moyennisation* ou *randomisation du jeu* (D, Θ, w).

Aux termes de la convention adoptée on écrira

$$\tilde{w}(\pi_{(\delta)}, \mathbf{Q}) = \tilde{w}(\delta, \mathbf{Q}),\ \tilde{w}(\pi, \mathbf{Q}_{(\theta)}) = \tilde{w}(\pi, \theta),$$

$$\tilde{w}(\delta, \theta) = w(\delta, \theta),$$

si $\pi_{(\delta)}$ et $\mathbf{Q}_{(\theta)}$ sont des distributions concentrées respectivement en δ et θ.

Il est clair que la moyennisation du jeu (D, Θ, w) exprime qu'on est passé à un jeu dont les ensembles de stratégies sont plus riches et par rapport auquel le couple initial est « plongé » : on obtiendrait ce jeu en considérant les seules stratégies pures des deux joueurs. Nous verrons que les problèmes d'ordonnancement des stratégies dans les jeux (D, Θ, w) et $(\tilde{D}, \tilde{\Theta}, \tilde{w})$ sont intimement liés.

DÉFINITION 4. On appelle *stratégie bayésienne associée à une distribution a priori* $\mathbf{Q}$ une stratégie $\pi = \pi_\mathbf{Q}$ telle que

$$\tilde{w}(\pi_\mathbf{Q}, \mathbf{Q}) = \inf_{\pi} \tilde{w}(\pi, \mathbf{Q}).$$

Une stratégie bayésienne n'est donc autre que la meilleure stratégie π pour **Q** donnée dans le jeu moyennisé.

On appelle *stratégie bayésienne pure* une stratégie $\delta_Q \in D$ telle que $\tilde{w}(\delta_Q, Q) = \inf_{\pi} \tilde{w}(\pi, Q)$.

THÉORÈME 1. *Si pour une distribution* **Q** *donnée, il existe une stratégie bayésienne mixte* π_Q, *il existera aussi une stratégie bayésienne pure* δ_Q *telle que*

$$\tilde{w}(\delta_Q, Q) = \tilde{w}(\pi_Q, Q).$$

DÉMONSTRATION. Elle coule presque de source. Désignons $a = \tilde{w}(\pi_Q, Q)$. Il est clair que

$$\tilde{w}(\delta, Q) \geqslant \inf_{\delta} \tilde{w}(\delta, Q) \geqslant a.$$

Si l'on admet que $\tilde{w}(\delta, Q) > a$ pour tous les δ, en prenant la moyenne par rapport à δ à l'aide de π_Q, on obtient

$$a = \int \tilde{w}(u, Q)\pi_Q(du) > a.$$

Cette contradiction prouve le théorème. ◄

Si donc $\inf_{\pi} \tilde{w}(\pi, Q)$ est atteint, il le sera sur les stratégies pures.

Si $\inf_{\delta} \tilde{w}(\delta, Q)$ n'est pas atteint, les stratégies bayésiennes n'existent pas et il est alors utile d'introduire la notion de *stratégie* ϵ*-bayésienne* qui existe toujours et qui se définit comme une stratégie δ_Q pour laquelle

$$\tilde{w}(\delta_Q, Q) \leqslant \inf_{\delta} \tilde{w}(\delta, Q) + \epsilon \tag{3}$$

pour $\varepsilon > 0$ donné. Mais, dans la suite, pour simplifier l'exposé on circonscrira notre étude aux problèmes dans lesquels les stratégies bayésiennes existent.

L'utilisation pratique des stratégies bayésiennes est un problème assez délicat. Si l'existence de la distribution *a priori* est conditionnée par un mécanisme physique réel, alors cette approche s'impose d'elle-même. Mais l'approche bayésienne a sa raison d'être dans les cas où elle est rattachée à des considérations éventuellement subjectives et pas toujours exhaustives qui doivent néanmoins être prises en considération. Le problème de l'utilisation de l'approche bayésienne sera examiné plus en détail dans le n° 4.

4. Stratégies minimax. Si l'on ne dispose pas d'une information *a priori* sur θ, on peut pour ordonner les stratégies partir de « la plus défavorable» stratégie de l'adversaire. Si l'on choisit une stratégie δ, les pertes maximales

seront

$$\sup_{\theta} w(\delta, \theta) \equiv w(\delta, \uparrow). \tag{4}$$

Cette quantité ne dépend que de δ et permet, de même que $w(\delta, Q)$ d'ordonner δ.

DÉFINITION 5. On dit qu'une stratégie $\bar{\delta}$ est *minimax* si

$$w(\bar{\delta}, \uparrow) = \inf_{\delta} w(\delta, \uparrow) \equiv w^*. \tag{5}$$

Le terme « minimax » tient son nom des opérations du second membre de la relation

$$w(\bar{\delta}, \uparrow) = \min_{\delta} \max_{\theta} w(\delta, \theta).$$

Il est évident que les stratégies minimax, tout comme les stratégies bayésiennes, peuvent en général ne pas exister. Dans ce cas, on peut introduire la notion de *stratégie ε-minimax* par analogie à (3). Dans la suite on admettra que sup et inf sont réalisés dans (4) et (5).

Etant donné que pour tout θ

$$w(\bar{\delta}, \theta) \leqslant w(\bar{\delta}, \uparrow) = w^*,$$

la stratégie minimax $\bar{\delta}$ fait subir au joueur 1 des pertes au plus égales à w^*.

DÉFINITION 6. Les valeurs

$$w^* = \inf_{\delta} w(\delta, \uparrow) \quad \left(w(\delta, \uparrow) = \sup_{\theta} w(\delta, \theta)\right),$$

$$w_* = \sup_{\theta} w(\downarrow, \delta) \quad \left(w(\downarrow, \delta) = \inf_{\delta} w(\delta, \theta)\right)$$

sont appelées respectivement *valeur supérieure* et *valeur inférieure du jeu*. Si $w^* = w_*$ on dit que *le jeu admet une valeur* qui est égale à la valeur commune w^* et w_*.

Il est clair de ce qui précède et pour des raisons de symétrie que si le joueur 2 adopte la même politique que le joueur 1 et choisit une stratégie minimax $\bar{\theta}$, il peut toujours s'assurer un gain $\geqslant w_*$. (Il aurait été plus correct d'appeler la stratégie $\bar{\theta}$ stratégie maximin, mais nous lui conserverons sa première appellation.) Donc, *si la valeur du jeu existe*, en choisissant une stratégie minimax $\bar{\delta}$, on s'assure un *résultat qui est inaméliorable* en ce sens que si l'adversaire opte pour $\bar{\theta}$, aucune autre stratégie ne nous causera des pertes inférieures à $w_* = w^*$. Il est évident que $w(\bar{\delta}, \bar{\theta}) = w^* = w_*$.

Dans le cas général, on a toujours $w^* \geqslant w_*$, puisque $w(\delta, \uparrow) \geqslant w(\delta, \theta) \geqslant w(\downarrow, \theta)$ pour tous les δ et θ et par suite

$$w^* = \inf_{\delta} w(\delta, \uparrow) \geqslant \sup_{\theta} w(\downarrow, \theta) = w_*. \tag{6}$$

Si $w^* > w_*$, *on peut améliorer* la stratégie minimax $\bar{\delta}$ en introduisant les stratégies mixtes. Ceci constitue l'un des principaux objectifs de ces dernières.

Désignons les stratégies minimax (si elles existent) du jeu moyennisé par $\bar{\pi}$ et $\bar{Q}$ respectivement et posons

$$\tilde{w}^* = \inf_{\pi} \sup_{Q} \tilde{w}(\pi, Q), \quad \tilde{w}_* = \sup_{Q} \inf_{\pi} \tilde{w}(\pi, Q).$$

Nous montrerons tout d'abord que les valeurs supérieure et inférieure d'un jeu se rapprochent par une moyennisation.

THÉORÈME 2. $w^* \geqslant \tilde{w}^* \geqslant \tilde{w}_* \geqslant w_*$.

DÉMONSTRATION. Elle est aussi élémentaire que celle du théorème 1. Vu que la randomisation du jeu peut être conduite en deux étapes : d'abord sur l'ensemble D et ensuite sur Θ, pour prouver ce théorème il suffit d'envisager seulement le moyennisé partiel $(\tilde{D}, \Theta, \tilde{w})$ du jeu (D, Θ, w). On a

$$\tilde{w}^* = \inf_{\pi} \sup_{\theta} \tilde{w}(\pi, \theta) \leqslant \inf_{\delta} \inf_{\theta} w(\delta, \theta) = w^*.$$

Puisque pour tous les π

$$\tilde{w}(\pi, \theta) = \int w(u, \theta)\pi(du) \geqslant \inf_{\delta} w(\delta, \theta) = w(\downarrow, \theta),$$

il vient $\inf_{\pi} \tilde{w}(\pi, \theta) \geqslant w(\downarrow, \theta)$,

$$\tilde{w}_* = \sup_{\theta} \inf_{\pi} \tilde{w}(\pi, \theta) \geqslant \sup_{\theta} w(\downarrow, \theta) = w_*.$$

L'inégalité $\tilde{w}^* \geqslant \tilde{w}_*$ a été prouvée dans (6). ◄

Le fait fondamental de la théorie des jeux est le théorème de minimax qui affirme que sous des conditions assez larges les *jeux moyennisés possèdent la valeur* $\tilde{w}^* = \tilde{w}_*$ *et admettent des stratégies minimax.*

Cette proposition sera formulée avec plus de rigueur dans le paragraphe prochain dans une situation plus générale pour des jeux statistiques.

Le jeu initial (D, Θ, w) ne possède généralement pas de valeur surtout si les ensembles D et Θ sont finis.

EXEMPLE 1. Considérons un jeu élémentaire dans lequel les ensembles D et Θ sont des doubletons : $D = \{\delta_1, \delta_2\}$, $\Theta = \{\theta_1, \theta_2\}$. Les valeurs de la fonction de perte $w(\delta, \theta)$ définissent une matrice $\|w(\delta_i, \theta_j)\|$, $i, j = 1, 2$, que nous poserons égale à $\begin{Vmatrix} 0 & 1 \\ 1 & 0 \end{Vmatrix}$. Ceci décrit un jeu dans lequel par exemple le joueur 1 doit deviner dans quelle main le joueur 2 a caché une pièce de monnaie. Une réponse juste fournit une perte nulle ($w(\delta_1, \theta_1) = w(\delta_2, \theta_2) = 0$), une réponse fausse, une perte de 1 rouble ($w(\delta_1, \theta_2) =$

$= w(\delta_2, \theta_1) = 1$). Il est évident que dans ce cas $w(\delta_i, \uparrow) = 1$, $w^* = 1$, $w(\downarrow, \theta_i) = 0$, $w_* = 0$, de sorte que le jeu ne possède pas de valeur et le joueur 1 ne peut pas s'assurer une perte inférieure à 1 rouble. La notion de stratégie minimax est inutile ici.

Moyennisons ce jeu. Les classes de stratégies $\tilde{D}$ et $\tilde{\Theta}$ sont ici les ensembles de toutes les distributions sur un doubleton. Il est évident que chacune des distributions sur D et Θ est décrite par la probabilité p et q de choix respectivement des stratégies δ_1 et θ_1. On peut donc admettre que $\tilde{D} = [0, 1]$, $\tilde{\Theta} = [0, 1]$. Les pertes du joueur 1 sont égales ici à

$$\tilde{w}(p, q) = p(1 - q) + q(1 - p) = p + q - 2pq,$$

$$\tilde{w}(p, \uparrow) = \begin{cases} p + 1 - 2p = 1 - p & \text{si } 2p < 1, \\ p & \text{si } 2p \geqslant 1. \end{cases}$$

$$\tilde{w}^* = 1/2.$$

On trouve de façon analogue que $\tilde{w}_* = 1/2$. Donc le jeu moyennisé possède dorénavant une valeur, et en choisissant les stratégies δ_1 et δ_2 avec une probabilité $p = 1 - p = 1/2$, le jouer 1 est assuré de perdre au plus 1/2. Cette stratégie ne peut être améliorée, car le joueur 2 s'assure le même gain en prenant $q = 1/2$.

Si un jeu moyennisé ne possède pas de valeur (cette situation ne peut se présenter que dans les jeux ayant un mécanisme complexe spécial), une deuxième moyennisation ne servirait à rien car elle sera pratiquement confondue avec la première.

Les approches bayésienne et minimax de résolution des problèmes de jeux sont largement répandues dans les activités humaines. L'approche bayésienne est basée sur le fait que l'on a une idée au moins approximative du comportement qu'adoptera le joueur 2. Le point de vue minimax se justifie dans les cas où il faut se prémunir contre des pertes élevées.

EXEMPLE 2. Un étudiant prépare un examen. On admettra qu'il n'a pas suffisamment de temps pour apprendre tout le programme. Par ailleurs son objectif est de décrocher la plus haute note possible.

Dans les conditions ci-dessus cet étudiant ne peut posséder à fond qu'une partie des sujets. Il se trouve donc devant l'alternative suivante : 1) apprendre sur le bout des doigts les seules questions qui sont le plus souvent posées par l'examinateur ; 2) connaître un peu de tout pour s'assurer une note passable. Le premier comportement correspondra à l'approche bayésienne, le second, à l'approche minimax.

Il est clair que la stratégie uniformément optimale consisterait ici à apprendre par cœur tout le programme, mais cette stratégie est exclue par hypothèse.

Dans certaines situations les stratégies minimax ne sont pas toujours les plus raisonnables.

EXEMPLE 3. Supposons que $\Theta = [0, 1]$ et que $D = \{\delta_1, \delta_2\}$. La fonction de perte est définie par les relations (fig. 9)

$$w(\delta_1, \theta) \equiv 1,$$
$$w(\delta_2, \theta) = 4(1 + \epsilon)\theta(1 - \theta).$$

Ici $w(\delta_1, \uparrow) = 1$, $w(\delta_2, \uparrow) = 1 + \epsilon$, $w^* = 1$, et la stratégie minimax sera δ_1, bien que la stratégie δ_2 soit la meilleure dans la « majorité » des cas pour

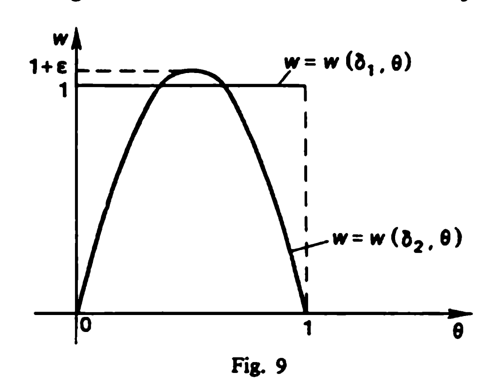


Fig. 9

de petits $\epsilon > 0$: $w(\delta_2,\theta) < 1$ si θ appartient au domaine $\left|\theta - \frac{1}{2}\right| > \frac{1}{2}\sqrt{\frac{\epsilon}{1 + \epsilon}}$. Les stratégies bayésiennes seront aussi confondues avec δ_2 pour la « majorité » des distributions $\mathbf{Q}$ sur $\Theta = [0, 1]$ (dont la masse n'est pas concentrée au voisinage du point $\theta = 1/2$).

Les notions de stratégies bayésienne et minimax sont liées entre elles. La proposition suivante nous fournit un procédé de recherche des stratégies minimax à l'aide de stratégies bayésiennes.

DÉFINITION 7. On dit qu'une stratégie $\bar{\pi}$ est une *stratégie niveleuse* sur un ensemble $\Theta_0 \subset \Theta$ si

1) $\tilde{w}(\bar{\pi}, \theta) = c = \text{const}, \theta \in \Theta_0$,

2) $\tilde{w}(\bar{\pi}, \theta) \leqslant c, \forall\theta$.

THÉORÈME 3. *Supposons qu'existent une distribution a priori $\bar{\mathbf{Q}}$ et la stratégie bayésienne correspondante $\pi_{\bar{Q}}$ qui est une stratégie niveleuse sur le support $N_{\bar{Q}}$ de la distribution $\bar{\mathbf{Q}}$. Alors $\bar{\pi} = \pi_{\bar{Q}}$ est une stratégie minimax.*

Si $N_{\bar{Q}} = \Theta$, la stratégie niveleuse $\bar{\pi}$ rend « indifférent » le jeu du joueur 2, c'est-à-dire indépendant de lui (comparer avec l'exemple 1).

DÉMONSTRATION du théorème 3. Désignons $\sup_\theta \tilde{w}(\pi, \theta) = \tilde{w}(\pi, \uparrow)$, $\inf_\delta \tilde{w}(\delta, Q) = \tilde{w}(\downarrow, Q)$. Nous devons nous assurer que

$$\tilde{w}(\pi_Q, \uparrow) = \inf_\pi \tilde{w}(\pi, \uparrow).$$

Ceci résulte des inégalités suivantes qui sont valables pour tout π :

$$\tilde{w}(\pi, \uparrow) \geqslant \tilde{w}(\pi, \bar{Q}) \geqslant \tilde{w}(\pi_Q, \bar{Q}) = \int \tilde{w}(\pi_Q, t)\bar{Q}(dt) = c \geqslant \tilde{w}(\pi_Q, \uparrow). \blacktriangleleft$$

L'extension suivante du théorème 3 est parfois utile.

THÉORÈME 3A. *Supposons qu'il existe des suites Q_n et π_{Q_n} telles que $\tilde{w}(\pi_{Q_n}, Q_n) \to c$. Supposons par ailleurs qu'il existe une stratégie $\bar{\pi}$ telle que $w(\bar{\pi}, \theta) \leqslant c$ pour tous les θ. Alors $\bar{\pi}$ est une stratégie minimax.*

DÉMONSTRATION. Elle est aussi élémentaire :

$$\tilde{w}(\pi, \uparrow) \geqslant \tilde{w}(\pi, Q_n) \geqslant \tilde{w}(\pi_{Q_n}, Q_n) \to c.$$

Ce qui n'est réalisé que si $\inf_\pi \tilde{w}(\pi, \uparrow) \geqslant c$. Comme $c \geqslant w(\bar{\pi}, \uparrow)$ le théorème est prouvé.

La distribution $\bar{Q}$ qui définit la stratégie bayésienne minimax π_Q dans le théorème 3 jouit de la remarquable propriété suivante : elle sera la plus défavorable en ce sens qu'elle maximise les pertes bayésiennes $\tilde{w}(\pi_Q, Q)$.

DÉFINITION 8. Une distribution $\bar{Q}$ est dite *la plus défavorable* ou *la moins bonne* si

$$\tilde{w}(\pi_Q, \bar{Q}) = \sup_Q \tilde{w}(\pi_Q, Q),$$

ou, en d'autres termes, si $\tilde{w}(\downarrow, \bar{Q}) = \sup_Q \tilde{w}(\downarrow, Q)$.

THÉORÈME 4. *Supposons que le jeu $(\tilde{D}, \bar{\Theta}, \tilde{w})$ possède une valeur et que les deux joueurs disposent des stratégies minimax $\bar{\pi}$ et $\bar{Q}$. La distribution $\bar{Q}$ est alors la plus défavorable et $\bar{\pi}$ est la stratégie bayésienne $\bar{\pi} = \pi_{\bar{Q}}$, qui est associée à $\bar{Q}$.*

REMARQUE 1. Le fait que le théorème 1 affirme l'existence conjointement à π_Q, d'une stratégie bayésienne pure δ_Q, ne signifie encore pas que cette dernière sera minimax aussi.

REMARQUE 2. D'après le théorème fondamental de minimax, la condition du théorème 4 relative à l'existence de la valeur du jeu moyennisé et des stratégies minimax ne doit pas être considérée comme une condition restrictive.

Nous aurons besoin de la proposition auxiliaire suivante que nous formulerons en termes du jeu initial (non moyennisé).

LEMME 1. *Supposons que le jeu (D, Θ, w) possède une valeur et que les stratégies minimax $\bar{\delta}$ et $\bar{\theta}$ des deux joueurs sont telles que*

$$w(\bar{\delta}, \uparrow) = \inf_{\delta} w(\delta, \uparrow), \quad w(\downarrow, \bar{\theta}) = \sup_{\theta} w(\downarrow, \theta).$$

Alors

$$w(\bar{\delta}, \uparrow) = w(\bar{\delta}, \bar{\theta}) = w(\downarrow, \bar{\theta}), \tag{7}$$

$$w^* = w(\bar{\delta}, \bar{\theta}) = w_*. \tag{8}$$

Réciproquement, si pour certaines stratégies $\bar{\delta}$, $\bar{\theta}$ la relation (7) *est remplie, alors la relation* (8) *le sera aussi et $\bar{\delta}$ et $\bar{\theta}$ seront des stratégies minimax.*

DÉMONSTRATION. Quels que soient δ et θ on a

$$w(\delta, \uparrow) \geqslant w(\delta, \theta) \geqslant w(\downarrow, \theta).$$

D'où

$$w^* = w(\bar{\delta}, \uparrow) \geqslant w(\bar{\delta}, \bar{\theta}) \geqslant w(\downarrow, \bar{\theta}) = w_*. \tag{9}$$

Vu que $w^* = w_*$ par hypothèse, tous les signes d'inégalité de (9) doivent être remplacés par des signes d'égalité. Ce qui prouve (7) et (8).

Réciproquement, si (7) a lieu, alors

$$w^* = \inf_{\delta} w(\delta, \uparrow) \leqslant w(\bar{\delta}, \uparrow) = w(\downarrow, \bar{\theta}) \leqslant \sup_{\theta} w(\downarrow, \theta) = w_*.$$

Vu que l'on a toujours $w^* \geqslant w_*$, les inégalités mentionnées expriment que $w^* = w_*$ et les stratégies $\bar{\delta}$ et $\bar{\theta}$ sont minimax. ◄

Le point $(\bar{\delta}, \bar{\theta})$ qui jouit de la propriété (7) s'appelle *point selle* ou *col* et le lemme 1, critère d'existence du col des stratégies minimax inaméliorables.

DÉMONSTRATION du théorème 4. Appliquons le lemme 1 au jeu moyennisé $(\tilde{D}, \tilde{\Theta}, \tilde{w})$. On trouve alors que

$$\tilde{w}(\bar{\pi}, \overline{\mathbf{Q}}) = \tilde{w}(\downarrow, \overline{\mathbf{Q}}) = \tilde{w}_* = \sup_{\mathbf{Q}} \tilde{w}(\downarrow, \mathbf{Q}).$$

D'où il résulte que la distribution $\overline{\mathbf{Q}}$ est la plus défavorable et que $\bar{\pi}$ est une stratégie bayésienne correspondant à $\overline{\mathbf{Q}}$. ◄

Les propositions exhibées plus haut peuvent désormais être résumées sous la forme du critère suivant de minimax qui décrit de façon complète le lien entre les stratégies minimax et les stratégies bayésiennes.

THÉORÈME 5. *Supposons que le jeu $(\tilde{D}, \tilde{\Theta}, \tilde{w})$ admet une valeur et des stratégies minimax. Les trois conditions suivantes sont alors équivalentes :*

1) *La stratégie $\bar{\pi}$ est minimax.*

2) *La stratégie $\bar{\pi}$ est bayésienne et niveleuse.*

3) *La stratégie $\bar{\pi}$ est bayésiene et correspond à la distribution la plus défavorable $\overline{\mathbf{Q}}$: $\bar{\pi} = \pi_{\overline{\mathbf{Q}}}$.*

DÉMONSTRATION. L'implication 2) ⇒ 1) a été prouvée dans le théorème 3 (la condition du théorème 5 n'est pas exigée pour cela). L'implication 1) ⇒ 3) a été établie dans le théorème 4. Reste à s'assurer que 3) ⇒ 2), c'est-à-dire que la stratégie bayésienne correspondant à la distribution la plus défavorable est niveleuse. On a

$$\tilde{w}_* = \tilde{w}(\bar{\pi}, \bar{Q}) = \int \tilde{w}(\bar{\pi}, t)\bar{Q}(dt) \leqslant \sup_t \tilde{w}(\bar{\pi}, t) = \tilde{w}^*.$$

Ce qui exprime que $\int \tilde{w}(\bar{\pi}, t)\bar{Q}(dt) = \sup_t \tilde{w}(\bar{\pi}, t)$ et par suite

$$\tilde{w}(\bar{\pi}, t) = \tilde{w}(\bar{\pi}, \uparrow) \quad [\bar{Q}]\text{-presque partout.}$$

Etant donné que d'autre part on a toujours $\tilde{w}(\bar{\pi}, t) \leqslant \tilde{w}(\bar{\pi}, \uparrow)$, il s'ensuit que $\bar{\pi}$ est une stratégie niveleuse. ◄

Revenons maintenant à l'application des classes de stratégies envisagées. Supposons que l'on n'ait pas réussi à trouver une sous-classe satisfaisante de stratégies contenant la stratégie uniformément optimale. Supposons par ailleurs que l'on se fait une certaine idée du comportement du joueur 2 (c'est-à-dire des valeurs prévisibles de θ) mais que ceci ne suffit pas pour appliquer l'approche bayésienne à son état pur. L'emploi de la démarche minimax dans ces conditions nous priverait de l'information existante. Dans une telle situation on peut se servir de l'approche intermédiaire suivante :

1. Tout d'abord il faut se prémunir contre les pertes élevées, c'est-à-dire n'envisager que les stratégies δ pour lesquelles $w(\delta, \theta) \leqslant w^* + a$ pour une valeur $a > 0$ convenable et quel que soit θ. L'ensemble des stratégies vérifiant cette inégalité sera désigné par D_a.

2. Dans ce sous-ensemble (c'est-à-dire dans le jeu (D_a, Θ, w)) on peut déjà appliquer l'approche bayésienne et utiliser les approximations qui nous sont accessibles pour la distribution *a priori* **Q**.

Cette approche mixte est constamment utilisée dans la vie courante. Dans les conditions de l'exemple 2, cette approche commande à l'étudiant d'apprendre un tout petit peu (pour éviter d'être recalé) tout le programme et ensuite de connaître un peu mieux les questions le plus souvent posées.

L'approche mixte doit comporter une étude mathématique de la stabilité des pertes bayésiennes dans le jeu (D_a, Θ, w) pour les variations permises de **Q**.

5. Classe complète de stratégies. Si les approches décrites ci-dessus ne permettent pas de déboucher sur une stratégie unique, on peut limiter la résolution du problème à la description de la classe complète des stratégies.

DÉFINITION 9. Une classe de stratégies $D^c \subset \tilde{D}$ est dite *complète* si pour toute stratégie $\pi \notin D^c$, il existe une stratégie $\pi_0 \in D^c$ qui est meilleure que π.

On dit qu'une classe complète D^c_0 est une *classe complète minimale* si aucune de ses sous-classes n'est complète.

En d'autres termes, une classe complète minimale est composée uniquement de stratégies inaméliorables.

L'utilité de construire une classe complète minimale ou une classe complète bien plus petite que D saute aux yeux. Cette procédure permet de réduire le jeu $(\tilde{D}, \tilde{\Theta}, \tilde{w})$ à un jeu $(D^c, \tilde{\Theta}, \tilde{w})$ de structure plus simple.

Le deuxième théorème fondamental de la théorie des jeux affirme que sous des hypothèses assez larges, *la classe de toutes les stratégies bayésiennes* $\{\pi_Q\}$, $Q \in \tilde{\Theta}$ *est une classe complète.* L'énoncé exact de ce théorème sera donné dans le paragraphe suivant. Dans certains cas les classes complètes peuvent être construites directement à l'aide de la structure du jeu. Supposons par exemple qu'il existe une partition de l'espace D en sous-ensembles D_b, $D = \bigcup_{b \in B} D_b$, $D_{b_1} \neq D_{b_2}$, $b_1 \neq b_2$, telle que chacun d'eux (c'est-à-dire pour les jeux (D_b, Θ, w)) contienne une stratégie uniformément optimale $\delta_b \in D_b$. Il est clair que dans ce cas la classe $D^c = \{\delta_b\}_{b \in B}$ sera complète. Ce point de vue sur la construction d'une classe complète sera illustré dans le § 3.

§ 3. Jeux statistiques

1. Description des jeux statistiques. Les principaux éléments d'un jeu statistique sont engendrés par le triplet (D, Θ, w) étudié dans le paragraphe précédent. Mais il faut leur ajouter ce qui suit :

1) Dans les jeux statistiques le joueur 1 est le *statisticien* (le chercheur), le joueur 2 est la *nature* (plus exactement la nature du phénomène étudié). Cette dernière choisit le paramètre (la stratégie) θ qui nous est inconnu et qui définit l'état de l'objet étudié. La plupart des problèmes de statistique mathématique sont liés d'une manière ou d'une autre à la prise de décisions δ qui « devinent » le plus exactement possible l'inconnue θ. Ceci étant, on aura présent à l'esprit que la nature n'aspire pas au gain maximal (c'est-à-dire n'aspire pas à nous causer le plus grand tort) et de ce point de vue est un joueur «indifférent » au choix de ses stratégies.

2) Dans les jeux statistiques, nous avons la possibilité de «sonder » la stratégie de la nature par des expériences qui nous fourniront sous forme d'un échantillon $X \in \mathbf{P}_\theta$ des « renseignements » sur la valeur éventuelle θ. Ainsi l'échantillon X de taille *) n distribué suivant la loi $\mathbf{P}_\theta$ qui dépend de θ est un élément du jeu statistique.

Dans ces conditions, la décision δ doit de toute évidence être choisie en fonction de X. Les stratégies du statisticien sont maintenant toutes les fonc-

*) Dans les constructions de ce paragraphe on aurait pu sans perdre en généralité admettre que $n = 1$. Mais nous envisagerons un échantillon de taille n pour conserver les liens élémentaires avec les résultats des chapitres précédents et des paragraphes 6, 7 et 8 suivants.

Une conception plus générale du jeu statistique fait intervenir un échantillon illimité $X_\infty = (x_1, x_2, \dots)$ dont chaque élément x_n est lié à des pertes $c_n \geqslant 0$ (cf. [54]).

tions $\delta(X)$ de $\mathscr{X}^n$ dans D. Les fonctions $\delta(X)$ s'appellent *fonctions de décision* ou *règles de décision* ou tout simplement *décisions*. On se bornera à étudier des applications *mesurables* $\delta(X)$ de $(\mathscr{X}^n, \mathfrak{B}^n_{\mathscr{X}})$ dans $(D, \mathfrak{F}_D)$. L'ensemble de ces applications sera désigné par $\mathscr{D}$.

L'ensemble Θ des stratégies du joueur 2 (la nature) reste le même.

Si nous choisissons la décision $\delta(X)$, et la nature, la valeur θ, nos pertes seront $w(\delta(X), \theta)$. On reconnaît ici une variable aléatoire. Cette situation n'est pas commode et pour l'éviter on prend naturellement pour pertes attachées à la stratégie $\delta = \delta(\cdot) \in \mathscr{D}$ et à $\theta \in \Theta$ la valeur de l'espérance mathématique

$$W(\delta(\cdot), \theta) = \mathbf{E}_\theta w(\delta(X), \theta) = \int w(\delta(x), \theta)\mathbf{P}_\theta(dx), \qquad (1)$$

qui porte le nom de *fonction de risque* (l'apparition du mot « risque » est logique, puisque l'utilisation de $\delta(\cdot)$ nous fournit un résultat aléatoire). Si la condition (A_μ), qui préconise que la distribution $\mathbf{P}_\theta$ admet une densité $f_\theta(x)$ par rapport à une mesure σ-finie μ, est remplie, la fonction de risque peut être mise sous la forme

$$W(\delta(\cdot), \theta) = \int w(\delta(x), \theta) f_\theta(x)\mu^n(dx).$$

Nous pouvons maintenant donner la définition suivante.

DÉFINITION 1. On appelle *jeu statistique* le triplet $(\mathscr{D}, \Theta, W)$, où Θ est l'ensemble des stratégies de la nature, $\mathscr{D}$, l'ensemble de toutes les applications mesurables de l'espace $\mathscr{X}^n$ dans l'ensemble D, W est définie dans (1). Pour caractériser plus complètement le jeu statistique, on peut donner en plus du triplet $(\mathscr{D}, \Theta, W)$ le couple $(X, \mathbf{P}_\theta)$, où $X \in \mathbf{P}_\theta$.

EXEMPLE 1. Supposons que $\theta \in [0, 1]$ définit le taux d'un composé chimique d'un minerai destiné à la fonte. Si nous décidons que le taux de ce composé est égal à $\delta \neq \theta$ et que la fonte soit conduite en fonction de cette décision, le métal obtenu sera de moindre qualité que pour $\delta = \theta$ et les dépenses d'énergie plus élevées. En d'autres termes, nous subirons des pertes $w(\delta, \theta)$ qui seront d'autant plus élevées que l'écart de δ à θ sera grand. Supposons pour simplifier que $w(\delta, \theta)$ est proportionnelle au carré de l'écart de δ à θ :

$$w(\delta, \theta) = c(\delta - \theta)^2.$$

(Si la fonction $w(\delta, \theta)$ est régulière et si l'on se place dans un voisinage de la droite $\delta = \theta$, la seule façon de simplifier le problème est d'admettre l'indépendance de c par rapport à θ.) On obtient alors un jeu (D, Θ, w) dans lequel $D = [0, 1]$ et $\Theta = [0, 1]$

$$w(\delta, \uparrow) = \sup_\theta w(\delta, \theta) = \begin{cases} c\delta^2 & \text{si } \delta > 1/2, \\ c(1 - \delta)^2 & \text{si } \delta \leqslant 1/2, \end{cases}$$

$$w^* = \inf_\delta w(\delta, \uparrow) = w(1/2, \uparrow) = c/4.$$

La stratégie $\delta = 1/2$ est donc minimax et garantit des pertes $\leqslant c/4$. Ce jeu est sans valeur, puisque $w_* = 0$. La moyennisation du jeu n'améliore pas la stratégie minimax $\delta = 1/2$ (elle fournit un $\tilde{w}_* = c/4$). Nous laissons au lecteur le soin de s'assurer que la stratégie bayésienne δ_Q est ici de la forme $\delta_Q = \mathbf{E}_Q\theta = \int t\mathbf{Q}(dt)$ (ceci résulte des égalités $\tilde{w}(\delta, \mathbf{Q}) = c\mathbf{E}_Q(\delta - \theta)^2 = c\mathbf{E}_Q(\theta - \mathbf{E}_Q\theta)^2 + c\mathbf{E}_Q(\delta - \mathbf{E}_Q\theta)^2)$ et que la distribution $\overline{\mathbf{Q}}$ la plus défavorable sera de la forme $\overline{\mathbf{Q}}(\{0\}) = \overline{\mathbf{Q}}(\{1\}) = 1/2$. Il est évident que la stratégie bayésienne correspondante est $\delta_{\overline{Q}} = 1/2$.

Supposons maintenant que le minerai n'est pas homogène et que l'on a la possibilité d'en analyser n échantillons. Ces échantillons sont prélevés de telle sorte que les résultats de l'analyse sont aléatoires et nous fournissent des valeurs indépendantes $(x_1, \ldots, x_n) = X$ à propos desquelles on sait que $\mathbf{E}x_i = \theta$, $\mathbf{V}x_i = b(\theta)$. Dans ce cas tous les estimateurs $\theta^* = \delta(X)$ du paramètre θ au vu de l'échantillon X seront les décisions $\delta(X)$. Le risque de la décision $\delta(X)$ sera égal à

$$W(\delta, \theta) = c\mathbf{E}_\theta(\delta(X) - \theta)^2,$$

et nous sommes conduits à la recherche de l'estimateur $\theta^* = \delta(X)$ minimisant ce risque dans un certain sens. Si l'on pose par exemple $\delta_1(X) = \bar{x}$, on obtient

$$W(\delta_1, \theta) = \frac{cb(\theta)}{n}. \tag{2}$$

Le maximum de $b(\theta)$ est égal à $\theta(1 - \theta)$ et il est atteint sur la distribution de x_1 concentrée en 0 et 1. Vu qu'il est possible d'exclure cette éventualité, on a

$$b(\theta) < \theta(1 - \theta) \leqslant 1/4, \quad W(\delta_1, \theta) < c/4n.$$

Donc, même pour $n = 1$ on obtient avec une stratégie que n'est pas éventuellement la meilleure un résultat meilleur qu'avec une stratégie minimax dans le jeu sans échantillon. La relation (2) montre également que le risque converge vers 0 lorsque $n \to \infty$. ◄

De la définition du jeu statistique, il ressort que le dernier jeu possède un ensemble $\mathscr{D}$ de stratégies qui est bien plus riche que pour le jeu initial (D, Θ, w).

Comme dans le § 2, conjointement au jeu $(\mathscr{D}, \Theta, W)$ dont les stratégies seront appelées *pures*, on peut envisager des *jeux randomisés* ou *mixtes* $(\tilde{\mathscr{D}}, \tilde{\Theta}, \tilde{W})$. L'ensemble $\tilde{\mathscr{D}}$ est celui des applications $\pi(X) : \mathscr{X}^n \to \tilde{D}$. Ces applications doivent être telles que les valeurs

$$\tilde{w}(\pi(X), \theta) = \int_D w(u, \theta)\pi(X, du)$$

soient aléatoires ($\pi(X, A)$ est la probabilité de l'ensemble $A \subset D$ conformément à la règle de décision π). On pose par définition

$$\tilde{W}(\pi(\cdot), \mathbf{Q}) = \int_{\Theta} \int_{\mathcal{X}^n} \int_{D} w(u, t)\pi(x, du)\mathbf{P}_t(dx)\mathbf{Q}(dt).$$

La stratégie $\pi(X)$ s'appelle *décision randomisée.*

Les relations d'ordre partiel sur les stratégies, les stratégies uniformément les meilleures, les stratégies bayésiennes et minimax, les classes complètes de stratégies pour les jeux statistiques se définissent exactement comme pour les jeux ordinaires (en remplaçant D par $\mathscr{D}$ et les fonctions w et $\tilde{w}$ par W et $\tilde{W}$).

Les théorèmes 2.1 à 2.5 se généralisent intégralement aux jeux statistiques, puisque ces derniers ne dépendent pas de la nature de l'ensemble D.

2. Classification des jeux statistiques. La classification suivante des jeux statistiques est liée à la nature des ensembles D et Θ :

1) Si $\Theta = A$, $D = A$, où A est un sous-ensemble « solide » de R^k (par exemple un parallélépipède), $w(t, t) = 0$, $w(t, u) > 0$ pour $t \neq u$, on obtient des problèmes d'estimation ponctuelle du paramètre inconnu θ.

2) Si les ensembles $\Theta = \{\theta_1, \ldots, \theta_r\}$ et $D = \{\delta_1, \ldots, \delta_r\}$ sont finis et contiennent le même nombre d'éléments, $w(\delta_i, \theta_i) = 0$, $w(\delta_i, \theta_j) > 0$ pour $i \neq j$, on obtient des problèmes de test d'un nombre fini d'hypothèses simples.

3) Si Θ est un domaine « solide » de R^k, $D = \{\delta_1, \delta_2\}$ est un doubleton, $w(\delta_1, \theta) = 0$ si $\theta \in \Theta_1$, $w(\delta_2, \theta) = 0$ si $\theta \in \Theta_2$ ($\Theta_1 \cap \Theta_2 = \varnothing$) et $w(\delta_i, \theta) > 0$ dans les autres cas, on est alors conduit à un problème de test des hypothèses $\{\theta \in \Theta_1\}$ et $\{\theta \in \Theta_2\}$.

Il existe certes d'autres classes de problèmes, mais nous avons distingué ces trois types parce qu'ils ont étés étudié dans les chapitres 2 et 3. Ces problèmes ont été envisagés d'un point de vue purement « statistique » impliquant un choix spécial des fonctions $w(\delta, \theta)$: les pertes ont été définies dans le premier groupe de problèmes par l'écart quadratique moyen, d'où la fonction de perte $w(\delta, \theta) = (\delta - \theta)^2$; dans le deuxième groupe, par la probabilité d'erreur, d'où la fonction de perte

$$w(\delta_i, \theta_j) = \begin{cases} 0, & i = j, \\ 1, & i \neq j. \end{cases}$$

Idem pour le troisième groupe dans lequel

$$w(\delta_1, \theta) = \begin{cases} 0 & \text{si} \quad \theta \in \Theta_1, \\ 1 & \text{si} \quad \theta \in \Theta_2, \end{cases}$$

$$w(\delta_2, \theta) = \begin{cases} 1 & \text{si} \quad \theta \in \Theta_1, \\ 0 & \text{si} \quad \theta \in \Theta_2. \end{cases}$$

Ces fonctions de perte qui correspondent au point de vue purement statistique seront dites *statistiques*.

Cette classification montre qu'il n'y a aucune différence de principe entre les problèmes de théorie de l'estimation et de théorie des tests d'hypothèses statistiques. Tout est dans la nature des ensembles Θ et D et dans la forme des fonctions de perte.

Profitons de cette classification pour signaler un trait spécifique des jeux statistiques (en complément aux n°s 1) et 2) de ce paragraphe) : l'ensemble D des jeux statistiques soit est confondu avec Θ, soit est un ensemble moins riche que Θ.

3. Deux théorèmes fondamentaux de théorie des jeux statistiques. Enonçons maintenant les résultats fondamentaux de la théorie des jeux statistiques. Nous avons déjà signalé que les théorèmes 2.1 à 2.5 se généralisaient aux jeux statistiques, car non liés à la nature de ces derniers. Pour établir les deux théorèmes fondamentaux mentionnés dans le § 2, nous aurons besoin de quelques conditions qui ne sont pas les plus générales (ce qui compliquerait énormément l'énoncé et la démonstration) mais qui sont tout de même assez larges pour englober les problèmes les plus intéressants et les plus profonds et, en particulier, les problèmes envisagés dans les chapitres 2 et 3.

CONDITION (A). *Chacun des ensembles Θ et D est ou bien fini ou bien un compact de R^k.*

Comme déjà signalé le cas où Θ est fini et $D \subset R^k$ peut ne pas être traité. Dans les trois autres cas, on admettra que la fonction de perte $w(\delta, \theta)$ satisfait la condition suivante.

CONDITION (B).

1) *Si $D \subset R^k$ et $\Theta \subset R^k$, la fonction $w(\delta, \theta)$ est continue sur $D \times \Theta$.*

2) *Si $\Theta \subset R^k$ et $D = \{\delta_1, \ldots, \delta_r\}$ est fini, chacune des r fonctions $w(\delta_i, \theta)$, $i = 1, \ldots, r$, est continue sur Θ.*

Si $\Theta = \{\theta_1, \ldots, \theta_r\}$ et $D = \{\delta_1, \ldots, \delta_r)$ sont finis, les fonctions $w(\delta_i, \theta_j)$, $i, j = 1, \ldots, r$, peuvent prendre des valeurs quelconques.

Nous exigerons de plus que soit remplie la

CONDITION (C). *Nous disposons d'un échantillon $X \in \mathbf{P}_\theta$ distribué suivant une loi $\mathbf{P}_\theta$ absolument continue pour tous les θ par rapport à une mesure σ-finie μ. Si $\Theta \subset R^k$, la densité $\frac{d\mathbf{P}_\theta}{d\mu}(x) = f_\theta(x)$ est continue dans $L_1(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mu)$ par rapport à θ, c'est-à-dire que lorsque $\theta_m \to \theta$*

$$\int |f_{\theta_m}(x) - f_\theta(x)|\mu(dx) \to 0. \tag{3}$$

Il est immédiat de vérifier que la continuité ordinaire de $f_\theta(x)$ par rapport à θ pour $[\mu]$-presque tous les x entraîne la continuité (3).

THÉORÈME 1. *Si les conditions* (A), (B) *et* (C) *sont réunies, le jeu moyen* $(\tilde{\mathfrak{D}}, \tilde{\Theta}, \tilde{W})$ *possède une valeur et des stratégies minimax* $\bar{\pi}(X)$ *et* $\overline{\mathbf{Q}}$:

$$\tilde{W}(\bar{\pi}(\cdot), \uparrow) = \inf_{\pi} \tilde{W}(\pi(\cdot), \uparrow), \quad \tilde{W}(\downarrow, \overline{\mathbf{Q}}) = \sup_{\mathbf{Q}} \tilde{W}(\downarrow, \mathbf{Q}).$$

Des théorèmes 2.4 et 2.5 du paragraphe précédent on sait que $\overline{\mathbf{Q}}$ est la distribution la plus défavorable,

$$\tilde{W}(\pi_{\mathbf{Q}}(\cdot), \overline{\mathbf{Q}}) = \sup_{\mathbf{Q}} \tilde{W}(\pi_{\mathbf{Q}}(\cdot), \mathbf{Q}) = \sup_{\mathbf{Q}} \tilde{W}(\downarrow, \mathbf{Q}),$$

et $\bar{\pi}(X) = \pi_{\mathbf{Q}}(X)$ est une stratégie bayésienne associée à $\overline{\mathbf{Q}}$.

On sait également qu'une condition nécessaire et suffisante pour que $\bar{\pi}(X)$ soit minimax (cf. théorème 2.5) est qu'elle soit bayésienne : $\bar{\pi}(X) = \pi_{\mathbf{Q}}(X)$ pour une distribution *a priori* $\mathbf{Q}$ et

$$\tilde{W}(\bar{\pi}(\cdot), \bar{\theta}) = c = \text{const} \quad \mathbf{Q}\text{-presque partout,}$$
$$\tilde{W}(\bar{\pi}(\cdot), \bar{\theta}) \leqslant c.$$

Ce critère minimax a été utilisé à maintes reprises et dans des situations particulières différentes (cf. §§ 2.11, 3.1, 3.5, 3.9).

THÉORÈME 2. *Si les conditions* (A), (B) *et* (C) *sont remplies, la classe de toutes les stratégies bayésiennes est complète.*

Dans l'annexe VIII, on démontre les théorèmes 1 et 2 dans leur forme plus générale où D et Θ sont des espaces métriques compacts (condition (A)) ; la fonction $w(\delta, \theta) : D \times \Theta \rightarrow R$ est continue par rapport à δ et θ pour les métriques correspondantes (condition (B)) ; la distribution $\mathbf{P}_\theta$ est continue par rapport à θ en variation (condition (C)).

Les démonstrations des théorèmes 1 et 2 sous certaines conditions accessoires sont accessibles dans [86]. Pour les cas où D et Θ sont finis, ces démonstrations figurent dans [7] et [89]. On peut trouver *ibidem* un exposé relativement complet des éléments de théorie générale des jeux statistiques (et en particulier une discussion de quelques cas de construction de la classe complète minimale ; cf. [89]).

Les théorèmes 1 et 2 soulignent toute l'importance de la description de la classe des décisions bayésiennes. Cette description fait l'objet du paragraphe suivant.

§ 4. Principe de Bayes. Classe complète de décisions

Nous savons que le jeu statistique est de par sa construction un objet plus compliqué que le jeu initial (D, Θ, w). Pour ce dernier la recherche des stratégies minimax et bayésiennes peut être relativement aisée notamment

dans le cas où les ensembles D et Θ sont de forme simple (par exemple finis). Dans le même temps, les jeux statistiques, même les plus élémentaires, présentent des ensembles $\mathscr{D}$ de nature assez complexe, ce qui complique sensiblement leur étude s'ils sont traités comme des jeux ordinaires.

EXEMPLE 1. Soient $D = \{\delta_1, \delta_2\}$, $\Theta = \{\theta_1, \theta_2\}$ des doubletons, $w(\delta_i, \theta_j) = w_{ij}$, $w_{ii} = 0$, $i, j = 1, 2$. Supposons que $\mathbf{Q} = (q, 1 - q)$ est une distribution *a priori* sur Θ. Alors

$$\tilde{w}(\delta_i, \mathbf{Q}) = qw_{i1} + (1 - q)w_{i2}.$$

Donc, la stratégie bayésienne $\pi_{\mathbf{Q}}$ est de la forme

$$\pi_{\mathbf{Q}}(\delta_2) = \begin{cases} 0 & \text{si} \quad \tilde{w}(\delta_1, \mathbf{Q}) < \tilde{w}(\delta_2, \mathbf{Q}) \quad (qw_{21} > (1 - q)w_{12}), \\ 1 & \text{si} \quad \tilde{w}(\delta_2, \mathbf{Q}) < \tilde{w}(\delta_1, \mathbf{Q}) \quad (qw_{21} < (1 - q)w_{12}) \end{cases} \tag{1}$$

($\pi_{\mathbf{Q}}(\delta_i)$ est la probabilité d'acceptation de δ_i).

Si

$$\tilde{w}(\delta_1, \mathbf{Q}) = \tilde{w}(\delta_2, \mathbf{Q}), \tag{2}$$

ou, ce qui est équivalent, $q = \bar{q} = w_{12}/(w_{12} + w_{21})$, on peut prendre pour $\pi_{\mathbf{Q}}$ n'importe quelle distribution π sur l'ensemble $\{\delta_1, \delta_2\}$. De façon exactement analogue, on peut toujours trouver une distribution $\pi = (p, 1 - p)$ telle que

$$\tilde{w}(\pi, \theta_1) = \tilde{w}(\pi, \theta_2) \quad \text{ou } pw_{12} = (1 - p)w_{21}.$$

La solution $\bar{p} = w_{21}/(w_{21} + w_{12})$ de cette équation correspond de toute évidence à la stratégie bayésienne niveleuse $\pi_{\bar{\mathbf{Q}}}$, $\bar{\mathbf{Q}} = (\bar{q}, 1 - \bar{q})$ qui sera minimax en vertu des théorèmes 2.4 et 2.5. La distribution $\bar{\mathbf{Q}}$ sera la plus défavorable.

Nous voyons que la « résolution » de ce problème est relativement simple. Si l'on passe au jeu statistique, même dans le cas élémentaire où $w_{12} = w_{21} = 1$, on obtient un problème sur les tests minimax et bayésiens dont l'étude a nécessité deux paragraphes : 3.1 et 3.2.

Le fait remarquable de ce paragraphe est que la recherche des stratégies bayésiennes (donc de la classe complète des stratégies minimax) pour les jeux statistiques peut être ramenée à celle des mêmes éléments pour les jeux initiaux (D, Θ, w). Cette réduction s'appuie sur la proposition suivante que nous appellerons *principe de Bayes*. Supposons comme précédemment que

$$f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i)$$

est la fonction de vraisemblance de l'échantillon X ; c'est aussi la densité de X dans $\mathscr{X}^n$ par rapport à μ^n. Supposons par ailleurs que la distribution *a priori* $\mathbf{Q}$ sur $(\Theta, \mathfrak{F}_\Theta)$ admet la densité $q(t)$ par rapport à une mesure λ (il

est évident que ce n'est pas une restriction). Alors, d'après le § 2.11, la fonction $f(x, t) = q(t)f_t(x)$ sera la densité de la distribution conjointe de (X, θ) dans $\mathscr{X}^n \times \Theta$. Ceci exprime que la fonction

$$q(t|x) = \frac{q(t)f_t(x)}{f(x)}, \qquad (3)$$

$$f(x) = \int q(t)f_t(x)\lambda(dt),$$

définit la densité conditionnelle de θ sachant que $X = x$. Cette densité correspond à la *distribution a posteriori* Q_x de la variable aléatoire θ sachant que $X = x$. La relation (3) s'appelle *formule de Bayes* (cf. §§ 2.10, 2.11).

THÉORÈME 1 (principe de Bayes). *Supposons que la condition* (A_μ) *est remplie, que la distribution a priori* Q *sur* Θ *admet* $q(t)$ *pour densité et que* Q_X *est la distribution a posteriori de densité* (3) *correspondant à* Q. *Supposons par ailleurs que le jeu initial* (D, Θ, w) *admet une stratégie bayésienne* π_Q *pour toute distribution a priori* Q. *Alors le jeu statistique* $(\mathscr{D}, \Theta, W)$ *admet une stratégie bayésienne* $\pi_Q(X)$ *correspondant à la distribution* Q *qui est confondue avec* π_{Q_x}, *stratégie bayésienne du jeu initial associée à la distribution a posteriori* Q_X.

La proposition de ce théorème peut être exprimée par une seule égalité :

$$\pi_Q(X) = \pi_{Q_x}.$$

Elle ramène le problème posé à la détermination de la distribution *a posteriori* Q_X et à la recherche des stratégies bayésiennes pour le jeu initial.

Le théorème 1 est capital pour l'appréhension du mécanisme de l'influence de l'information extraite de l'échantillon sur le choix de la stratégie optimale. L'information *a priori* fournie par la distribution Q sur Θ est constamment modifiée par les données expérimentales. La stratégie optimale sera celle qui tient compte de cette modification de la manière suivante : il faut prendre la stratégie optimale du jeu initial qui correspond non plus à Q mais à Q_X.

DÉMONSTRATION du théorème 1. On a

$$\tilde{W}(\pi(\cdot), Q) = \int_\Theta \int_{\mathscr{X}^n} \tilde{w}(\pi(x), t)f_t(x)\mu^n(dx)q(t)\lambda(dt) =$$

$$= \int_{\mathscr{X}^n} f(x)\mu^n(dx) \int_\Theta \tilde{w}(\pi(x), t)q(t|x)\lambda(dt). \qquad (4)$$

On s'est servi de (3). Le changement d'ordre d'intégration est licite, puisque l'intégrant est une fonction positive. La deuxième intégrale du dernier mem-

bre de (4) n'est autre que $\tilde{w}(\lambda(x), Q_x)$. Or pour tout x, on a

$$\tilde{w}(\pi(x), Q_x) \geqslant \tilde{w}(\pi_{Q_x}, Q_x) = \int_\Theta \tilde{w}(\pi_{Q_x}, t)q(t|x)\lambda(dt).$$

En portant cette inégalité dans (4) et en revenant à l'ordre initial d'intégration, on obtient

$$\tilde{W}(\pi(\cdot), Q) \geqslant \int_{\mathscr{X}^n} f(x)\mu^n(dx)\int_\Theta \tilde{w}(\pi_{Q_x}, t)q(t|x)\lambda(dt) = \tilde{W}(\pi_Q, Q).$$

Ce qui exprime, puisque $\pi(x)$ est arbitraire ici, que

$$\pi_Q(x) = \pi_{Q_x}. \blacktriangleleft$$

REMARQUE 1. En toute rigueur on devrait conjecturer la mesurabilité de la fonction $\tilde{w}(\pi_{Q_x}, t)$ par rapport à $\mathfrak{B}^n \times \mathfrak{F}_\Theta$. Mais on peut lever cette restriction dans la mesure où elle revêt un caractère purement technique et est superflue lorsque les conditions (A), (B), (C) du § 3 sont remplies. Le lecteur peut vérifier seul la dernière assertion en se servant du fait que si D et Θ sont discrets, cette mesurabilité s'établit de façon évidente, et que si les conditions (A) et (B) sont réunies, un jeu arbitraire peut être « approché » d'aussi près que l'on veut par un jeu discret.

Si l'on retourne à l'exemple 1, on peut maintenant s'appuyer sur le théorème 1 et indiquer aussitôt la forme des stratégies bayésiennes pour le jeu statistique correspondant. Plus exactement, de (1) on déduit

$$\pi_{Q_X}(\delta_2) = \begin{cases} 0 & \text{si} \quad q_X \equiv \dfrac{qf_{\theta_1}(X)}{qf_{\theta_1}(X) + (1-q)f_{\theta_2}(X)} > \dfrac{w_{12}}{w_{12}+w_{21}}, \\ 1 & \text{si} \quad q_X < \dfrac{w_{12}}{w_{12}+w_{21}}. \end{cases} \tag{5}$$

Si

$$q_X = \frac{w_{12}}{w_{12}+w_{21}}, \tag{6}$$

pour π_{Q_X} on peut prendre n'importe quelle distribution sur $\{\delta_1, \delta_2\}$. L'inégalité (5) peut alors s'écrire

$$\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} > \frac{a(1-q)}{q(1-a)}, \quad a = \frac{w_{12}}{w_{12}+w_{21}}. \tag{7}$$

On reconnaît ici le test du rapport de vraisemblance.

Par ailleurs

$$\tilde{W}(\pi_{Q\cdot}, \theta_j) = w_{1j}\mathsf{E}_{\theta_j}\pi_{Q_X}(\delta_1) + w_{2j}\mathsf{E}_{\theta_j}\pi_{Q_X}(\delta_2), \; j = 1, 2.$$

Supposons pour simplifier que l'égalité (6) est réalisée avec une $\mathbf{P}_{\theta_j}$-probabilité nulle, de sorte que la stratégie bayésienne sera pure avec une $\mathbf{P}_{\theta_j}$-probabilité égale à 1, $j = 1, 2$. Alors

$$\mathbf{E}_{\theta_j}\pi_{\mathbf{Q}X}(\delta_1) = \mathbf{P}_{\theta_j}\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} > \frac{a(1-q)}{q(1-a)}\right),$$

$$\tilde{W}(\pi_{\mathbf{Q}.}, \theta_1) = w_{21}\mathbf{P}_{\theta_1}\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} < \frac{a(1-q)}{q(1-a)}\right),$$

$$\tilde{W}(\pi_{\mathbf{Q}.}, \theta_2) = w_{12}\mathbf{P}_{\theta_2}\left(\frac{f_{\theta_1}(X)}{f_{\theta_2}(X)} > \frac{a(1-q)}{q(1-a)}\right).$$

De là on déduit sans peine la valeur $\bar{q}$ correspondant à la distribution la plus défavorable $\overline{\mathbf{Q}}$ pour laquelle la stratégie $\pi_{\mathbf{Q}X}$ sera niveleuse, c'est-à-dire telle que

$$\tilde{W}(\pi_{\mathbf{Q}.}, \theta_1) = \tilde{W}(\pi_{\mathbf{Q}.}, \theta_2).$$

Cette stratégie sera minimax en vertu des théorèmes 2.4 et 2.5.

Nous laissons au lecteur le soin de généraliser cette procédure de recherche d'une stratégie minimax au cas où les $\mathbf{P}_{\theta_1}$- ou $\mathbf{P}_{\theta_2}$-distributions de $f_{\theta_1}(X)/f_{\theta_2}(X)$ contiennent une composante discrète.

En s'appuyant sur le théorème 1, on peut de façon analogue généraliser les résultats des §§ 3.1 et 3.2 au cas d'ensembles D et Θ finis arbitraires et d'une fonction de perte arbitraire $w(\delta_i, \theta_j) = w_{ij}$ que l'on pourra appeler aussi matrice des pertes $\| w(\delta_i, \theta_j) \|$. (Dans les §§ 3.1 et 3.2, nous avons envisagé le cas particulier $w_{ij} = 1$ pour $i \neq j$.) Si les w_{ij} sont arbitraires, la décision bayésienne sera de la forme suivante. Supposons que $\mathbf{Q} = (q(\theta_1), \ldots$ $\ldots, q(\theta_r))$, $\mathbf{Q}_X = (q_X(\theta_1), \ldots, q_X(\theta_r))$,

$$q_X(\theta_j) = \frac{q(\theta_j)f_{\theta_j}(X)}{\sum_i q(\theta_i)f_{\theta_i}(X)}.$$

Alors $\tilde{w}(\delta_i, \mathbf{Q}_X) = \sum_{j=1}^{r} w_{ij}q_X(\theta_j)$ et par suite

$$\pi_{\mathbf{Q}X}(\delta_k) = 1 \text{ si } \tilde{w}(\delta_k, \mathbf{Q}_X) \leqslant \tilde{w}(\delta_i, \mathbf{Q}_X), \ \forall i,$$

ou ce qui est équivalent si

$$\sum_{j=1}^{r} w_{kj}f_{\theta_j}(X)q(\theta_j) \leqslant \sum_{j=1}^{r} w_{ij}f_{\theta_j}(X)q(\theta_j).$$

S'il existe quelques valeurs k (que l'on désignera par $k_1, \ldots, k_s$) jouissant de cette propriété, pour stratégie bayésienne $\pi_{\mathbf{Q}X}$ on peut prendre n'importe quelle distribution sur $\{\delta_{k_1}, \ldots, \delta_{k_s}\}$.

La recherche d'une stratégie minimax se déroule comme suit. Supposons, toujours par souci de simplicité, que les $\mathbf{P}_{\theta_j}$-distributions de $\tilde{w}(\delta_i, \mathbf{Q}_X)$ ne possèdent pas de composantes discrètes. Alors

$$\tilde{W}(\pi_{\mathbf{Q}_\cdot}, \theta_j) = \sum_{i \neq j} w_{ij} \mathbf{P}_{\theta_j}\left(\tilde{w}(\delta_i, \mathbf{Q}_X) < \min_{l \neq i} \tilde{w}(\delta_l, \mathbf{Q}_X)\right).$$

Le théorème 3.1 affirme qu'il existe une distribution $\overline{\mathbf{Q}} = (\bar{q}(\theta_1), \ldots \ldots, \bar{q}(\theta_r))$ pour laquelle la stratégie $\pi_{\overline{\mathbf{Q}}X}$ nivellera les valeurs de $\tilde{W}(\pi_{\overline{\mathbf{Q}}X}, \theta_j)$ pour tous les j. Cette stratégie sera justement une stratégie minimax.

En s'appuyant sur les raisonnements précédents et sur le théorème 3.2, on détermine aussi sans peine la forme de la classe complète des stratégies du jeu statistique $(\mathscr{D}, \Theta, \tilde{W})$ lorsque D et Θ sont finis.

Considérons les stratégies $\pi_{\mathbf{Q}X}$ qui sont les distributions sur les $\delta_{k_1}, \ldots \ldots, \delta_{k_s}$ pour lesquelles

$$\min_i \left(\sum_{j=1}^{r} (w_{k_l, j} - w_{ij}) f_{\theta_j}(X) q(\theta_j) \right) = 0.$$

La classe de ces stratégies (bayésiennes), obtenue en faisant prendre à $q(\theta_1), \ldots, q(\theta_r)$ toutes les valeurs possibles, sera une classe complète. Nous avons vu que pour $r = 2$, cette classe est très simple et étroite (cf. (7)) : elle est constituée des décisions $\pi(X) = (\pi(X, \delta_1), \pi(X, \delta_2))$, où $\pi(X, \delta_i)$ sont les probabilités d'acceptation de la décision δ_i

$$\pi(X, \delta_1) = \begin{cases} 1 & \text{si} \quad R(X) > c, \\ p \in [0, 1] & \text{si} \quad R(X) = c, \\ 0 & \text{si} \quad R(X) < c, \end{cases}$$
$$R(X) = \frac{f_{\theta_1}(X)}{f_{\theta_2}(X)}, \ 0 \leqslant c \leqslant \infty. \tag{8}$$

Dans les jeux où les ensembles D et Θ ont la puissance du continu, on peut trouver les décisions sous leur forme explicite pour certaines fonctions de pertes importantes. Supposons par exemple que D et Θ sont des domaines de R^k et que la fonction de perte est quadratique :

$$w(\delta, \theta) = c|\delta - \theta|^2 = c\sum_{i=1}^{k} |\delta_i - \theta_i|^2, \tag{9}$$

où δ_i et θ_i sont les coordonnées de δ et de θ. Alors

$$\tilde{w}(\delta, \mathbf{Q}) = c\int |\delta - t|^2 \mathbf{Q}(dt) = c\mathbf{E}_{\mathbf{Q}}|\delta - \theta|^2.$$

Nous savons que cette expression atteint son minimum pour $\delta = \mathbf{E}_{\mathbf{Q}}\theta = \int t\mathbf{Q}(dt)$. Ceci est visiblement la stratégie bayésienne $\delta_{\mathbf{Q}} = \mathbf{E}_{\mathbf{Q}}\theta$. De là et du principe de Bayes, il résulte que dans un jeu statistique la stratégie bayé-

sienne $\delta_Q(X) = \theta_Q^*$ sera de la forme

$$\theta_Q^* = \delta_{Q_X} = \int_{R^k} t Q_X(dt) = \int_{R^k} tq(t|X)\lambda(dt). \qquad (10)$$

Ce résultat a déjà été établi dans le chapitre 2.

Le risque de la stratégie θ_Q^* est égal à $W(\theta_Q^*, \theta) = c\mathbf{E}_\theta|\theta_Q^* - \theta|^2$. La distribution *a priori* $\bar{Q}$ pour laquelle $\mathbf{E}_\theta|\theta_Q^* - \theta|^2 = \text{const}$ nous donne l'estimateur minimax $\theta^* = \delta_{\bar{Q}}(X)$. On trouvera dans le § 2.11 des exemples de construction d'estimateurs minimax.

La classe des estimateurs (10), où Q parcourt toutes les distributions sur Θ, est une classe complète.

Considérons maintenant un autre cas particulier de fonction de perte

$$w(\delta, \theta) = c|\delta - \theta| \qquad (11)$$

et supposons que $\Theta = R$, $D = R$. Alors

$$\tilde{w}(\delta, Q) = c\mathbf{E}_Q|\delta - \theta| = c\int|\delta - t|Q(dt) =$$
$$= c\int_{-\infty}^{\delta}(\delta - t)Q(dt) + c\int_{\delta}^{\infty}(t - \delta)Q(dt).$$

En intégrant par parties et en posant $F(t) = Q(]-\infty, t[)$, on trouve

$$\tilde{w}(\delta, Q) = c\int_{-\infty}^{\delta}(\delta - t)dF(t) - c\int_{\delta}^{\infty}(t - \delta)d(1 - F(t)) =$$
$$= c\left[\int_{-\infty}^{\delta}F(t)dt + \int_{\delta}^{\infty}(1 - F(t))dt\right].$$

La dérivée de cette expression par rapport à δ existe presque partout et vaut $c[2F(\delta) - 1]$. Cette fonction est monotone croissante et change de signe au point $\bar{\delta}$ qui est égal à la médiane de la distribution $F: F(\bar{\delta} - 0) \leqslant 1/2$, $F(\bar{\delta} + 0) \geqslant 1/2$. D'où il résulte que $\tilde{w}(\delta, Q)$ est convexe par rapport à δ et présente un minimum au point $\bar{\delta}$.

D'après le principe de Bayes, ceci exprime que *la médiane de la distribution a posteriori* Q_X *sera l'estimateur bayésien* $\theta_Q^* = \delta_Q(X)$ *pour la distribution a priori* Q *et la fonction de perte* (11). Ceci nous permet, comme dans le cas (9), de trouver la décision minimax et la classe complète.

On pourrait traiter de façon analogue le cas

$$w(\delta, \theta) = c|\delta - \theta|^\alpha, \ \alpha > 0.$$

Signalons en conclusion de ce paragraphe que la fonction de perte quadratique (9) pour $c = 1$ et des ensembles D et Θ ayant la puissance du continu, et la fonction de perte

$$w(\delta_i, \theta_j) = \begin{cases} 0, & i = j, \\ 1, & i \neq j, \end{cases} \qquad (12)$$

pour D et Θ finis occupent une place particulière en théorie des jeux statistiques. Les fonctions de risque correspondantes se transforment respectivement en la somme de la variance et du carré du biais de l'estimateur si D et Θ ont la puissance du continu et en la probabilité de se tromper, si D et Θ sont finis. Ces caractéristiques qui sont naturelles en soi nous ont servi de base pour choisir les décisions optimales dans les chapitres 2, 3 et 4. Si le jeu statistique ne contient aucune indication quant à la forme de la fonction $w(\delta, \theta)$, alors le plus souvent on prend pour telle la fonction (9) ou la fonction (12). Nous avons convenu de les appeler *fonctions de perte statistiques.*

§ 5. Exhaustivité, absence de biais, invariance

Les principes d'exhaustivité, d'absence de biais et d'invariance nous servent à restreindre la classe des décisions. Ces principes nous commandent de prendre pour décisions uniquement des décisions respectivement exhaustives, sans biais et invariantes. L'utilisation de l'un, de deux ou des trois principes (quand cela est possible) nous permet dans bien des cas de restreindre la classe des stratégies envisagées à un point tel que son intersection avec la classe complète comprend une seule décision. Ceci exprime que la sous-classe ainsi définie contiendra la stratégie uniformément la meilleure (comparer avec le n°1 du § 2), donc que le problème de choix d'une décision est résolu.

Ces principes sont assez naturels et nous les avons discutés dans divers problèmes concrets dans les chapitres 2 et 3.

Le plus indiscutable d'entre eux est le principe d'exhaustivité, principe qui n'est souvent qu'un procédé de description d'une classe complète.

1. Exhaustivité. Supposons qu'est remplie la condition (A_μ) et qu'il existe une statistique exhaustive S, c'est-à-dire (cf. § 2.12)

$$f_\theta(X) = \psi(\theta, S) \cdot h(X).$$

Supposons par ailleurs qu'une distribution *a priori* $\mathbf{Q}$ possède une densité $q(t)$ par rapport à une mesure λ. Alors en vertu du principe de Bayes la stratégie bayésienne sera entièrement définie par la densité *a posteriori*

$$q(t|X) = \frac{q(t)f_t(X)}{\int q(u)f_u(X)\lambda(du)} = \frac{q(t)\psi(t, S)}{\int q(u)\psi(u, S)\lambda(du)}, \qquad (1)$$

qui dépend uniquement de S. Vu que toute distribution $\mathbf{Q}$ admet une densité par rapport à une mesure λ convenablement choisie (on peut par exemple poser $\lambda = \mathbf{Q}$, $q(t) \equiv 1$), ce qui vient d'être dit exprime que toutes les décisions bayésiennes $\pi_{\mathbf{Q}}(X)$ seront des fonctions uniquement de S :

$$\pi_{\mathbf{Q}}(X) = p_{\mathbf{Q}}(S).$$

En d'autres termes, aucune stratégie bayésienne $\pi_Q(X)$ ne dépend de X pour S fixe.

Supposons maintenant que sont remplies les conditions (A), (B) et (C) du § 3. La proposition ci-dessus sera valable aussi pour les stratégies minimax. Elle exprimera également que toutes les décisions qui sont fonctions uniquement de S (c'est-à-dire toutes les applications mesurables de $\mathbb{S} \to \bar{D}$, où $\mathbb{S}$ est l'ensemble des valeurs de S) forment une classe complète $\mathscr{D}_{\mathbb{S}}$. Ceci résulte du fait que $\mathscr{D}_{\mathbb{S}}$ contient toutes les stratégies bayésiennes qui, on le sait, forment une classe complète. Il est évident que la classe $\mathscr{D}_{\mathbb{S}}$ sera la plus petite pour la statistique exhaustive minimale S.

Il est clair que la classe complète minimale ne renferme pas toutes les fonctions de S (à valeurs dans $\bar{D}$), mais une faible partie seulement. Ce fait est corroboré par la formule (1) d'où il s'ensuit, par exemple, que pour les doubletons D et Θ (cf. (4.8)) la classe complète est constituée des fonctions $\pi(X)$ pour lesquelles la probabilité $\pi(X, \delta_1)$ d'acceptation de la décision δ_1 a la forme de l'indicateur de l'ensemble $\{R(X) > c\}$, où $R(X) = \psi(\theta_1, S)/\psi(\theta_2, S)$ (pour plus de précision voir (4.8)).

Si $D \subset R^k$, $\Theta \subset R^k$, et la fonction de perte $w(\delta, \theta)$ est de la forme $w(\delta, \theta) = w(\delta - \theta)$, où $w(u)$ est une fonction convexe dans R^k, on peut conférer au principe d'exhaustivité une forme constructive qui permet de caractériser efficacement la classe complète. Plus exactement, on a la généralisation suivante du théorème 2.14.1.

THÉORÈME 1 (Blackwell). *Pour toute décision (estimateur)* $\theta^* = \delta(X)$, *il existe un estimateur*

$$\theta^*_S = \mathbf{E}_\theta(\theta^* | S)$$

(θ^*_S est indépendant de θ, puisque S est exhaustive) qui est aussi bon que θ^*. Plus exactement, pour tout $\theta \in \Theta$

$$\mathbf{E}_\theta w(\theta^*_S - \theta) \leqslant \mathbf{E}_\theta w(\theta^* - \theta).$$

DÉMONSTRATION. On a l'inégalité de Jensen suivante (cf. § 2.9) : si g est une fonction convexe dans R^k, ξ une variable aléatoire à valeurs dans R^k et $\mathfrak{F}$ une sous-tribu de la tribu principale, alors

$$\mathbf{E}(g(\xi)|\mathfrak{F}) \geqslant g(\mathbf{E}(\xi|\mathfrak{F})).$$

Cette inégalité entraîne

$$\mathbf{E}_\theta w(\theta^* - \theta) = \mathbf{E}_\theta\{\mathbf{E}_\theta(w(\theta^* - \theta)|S)\} \geqslant$$
$$\geqslant \mathbf{E}_\theta w(\mathbf{E}_\theta(\theta^* - \theta|S)) = \mathbf{E}_\theta w(\theta^*_S - \theta). \quad ◄$$

Si la statistique exhaustive S est complète, le théorème 1 combiné au principe d'absence de biais nous permet de définir de façon unique l'estima-

teur uniformément le meilleur. En effet, soit K_0 la classe des estimateurs sans biais $\theta^* = \delta(X)$:

$$\mathbf{E}_\theta \theta^* = \theta \quad \text{pour} \quad \theta^* \in K_0.$$

En reprenant *ad litteram* les raisonnements du § 2.14 (théorème 3) on s'assure alors que $\theta^*_S = \mathbf{E}_S(\theta^*|S)$ sont confondus pour tous les $\theta^* \in K_0$ et, par suite, l'intersection de K_0 et de la classe complète est composée d'un seul estimateur de $\varphi(S)$ qui sera naturellement appelé *efficace*.

De ce qui précède il est évident que *les estimateurs efficaces, s'ils existent, seront les mêmes pour toute fonction de perte convexe* $w(\delta - \theta)$. Ceci permet d'appliquer à une telle fonction de perte tous les théorèmes respectifs établis dans le chapitre 2 pour $w(u) = u^2$.

Ces raisonnements montrent tout le parti que l'on peut tirer de l'application simultanée des principes d'exhaustivité et d'absence de biais.

2. Absence de biais. Nous venons tout juste de voir quel rôle peut jouer le principe d'absence de biais en théorie de l'estimation. Au § 3.6 on a montré que l'on pouvait obtenir le même effet (l'existence des tests sans biais uniformément les plus puissants) en utilisant les tests sans biais dans la théorie de tests d'hypothèses statistiques.

Dans le cas général l'absence de biais se définit comme suit. Supposons que le problème de décision consiste à « déterminer » la valeur inconnue θ et que par conséquent les ensembles D et Θ sont confondus. La fonction de perte $w(\delta, \theta)$ peut être arbitraire.

DÉFINITION 1. On dit qu'une décision $\delta(X)$ est *sans biais* si

$$\mathbf{E}_\theta w(\delta(X), \theta) \leqslant \mathbf{E}_\theta w(\delta(X), \theta')$$

quels que soient θ, $\theta' \neq \theta$.

En d'autres termes, $\min\limits_v \mathbf{E}_\theta w(\delta(X), v)$ est réalisé pour $v = \theta$. Ceci exprime que $\delta(X)$ se trouve en moyenne plus près de l'inconnue θ que de tout autre point.

Il est immédiat de voir que la définition antérieure de l'absence de biais est un cas particulier de celle-ci.

Si l'on teste deux hypothèses multiples $H_1 = \{\theta \in \Theta_1\}$ et $H_2 = \{\theta \in \Theta_2\}$, l'ensemble $D = \{\delta_1, \delta_2\}$ peut être fondamentalement différent de Θ. Dans ce cas la définition de l'absence de biais sera formellement différente, bien que sa signification soit la même. Plus exactement, la définition 1 peut se transformer (cf. [50]) en la

DÉFINITION 1A. On dit qu'une décision $\delta(X)$ est *sans biais* si

$$\mathbf{E}_\theta w(\delta(X), \theta) \leqslant \mathbf{E}_\theta w(\delta(X), \theta')$$

quels que soient $\theta \in \Theta_1$, $\theta' \in \Theta_2$ ou $\theta \in \Theta_2$, $\theta' \in \Theta_1$.

Supposons pour simplifier que

$$w(\delta_1, \theta) = w_1 = \text{const pour } \theta \in \Theta_2 \ ;$$
$$w(\delta_2, \theta) = w_2 = \text{const pour } \theta \in \Theta_1 \ ,$$

$\delta_1 = 0$, $\delta_2 = 1$ et $\delta(X)$ est la probabilité (égale à 1 ou 0) d'acceptation de H_2. Alors

$$\mathbf{E}_\theta w(\delta(X), \theta) = \begin{cases} w_2 \mathbf{P}_\theta(\delta(X) = 1) & \text{si } \theta \in \Theta_1, \\ w_1 \mathbf{P}_\theta(\delta(X) = 0) & \text{si } \theta \in \Theta_2, \end{cases}$$

$$\mathbf{E}_\theta w(\delta(X), \theta') = \begin{cases} w_1 \mathbf{P}_\theta(\delta(X) = 0) & \text{si } \theta \in \Theta_1, \theta' \in \Theta_2, \\ w_2 \mathbf{P}_\theta(\delta(X) = 1) & \text{si } \theta \in \Theta_2, \theta' \in \Theta_1. \end{cases}$$

et l'inégalité de la définition 1A exprime que

$$w_2 \mathbf{P}_{\theta_1}(\delta(X) = 1) \leqslant w_1 \mathbf{P}_{\theta_1}(\delta(X) = 0) \quad \text{si } \theta_1 \in \Theta_1,$$
$$w_1 \mathbf{P}_{\theta_2}(\delta(X) = 0) \leqslant w_2 \mathbf{P}_{\theta_2}(\delta(X) = 1) \quad \text{si } \theta_2 \in \Theta_2,$$

ou ce qui est équivalent

$$\mathbf{P}_{\theta_1}(\delta(X) = 1) \leqslant \frac{w_1}{w_1 + w_2}, \quad \mathbf{P}_{\theta_2}(\delta(X) = 1) \geqslant \frac{w_1}{w_1 + w_2}.$$

D'où il s'ensuit

$$\sup_{\theta \in \Theta_1} \mathbf{E}_\theta \delta(X \leqslant \inf_{\theta \in \Theta_2} \mathbf{E}_\theta \delta(X),$$

donc le test δ sera sans biais au sens de la définition du § 3.6. Réciproquement, si la dernière inégalité a lieu, le test δ sera sans biais au sens de la définition 1A pour une fonction de perte $w(\delta, \theta)$ convenablement choisie, par exemple, pour $w_1/(w_1 + w_2) = \sup_{\theta \in \Theta_1} \mathbf{E}_\theta \delta(X)$.

On trouvera d'autres exemples d'application du principe d'absence de biais (en plus des résultats du § 3.6) dans [50].

3. **Invariance.** Nous avons vu que l'intersection de la classe complète engendrée par les décisions « exhaustives » avec la classe des décisions sans biais pouvait être composée d'une seule stratégie. Une autre classe naturelle de stratégies susceptible de contenir une seule décision inaméliorable est la classe des décisions invariantes (comparer avec les §§ 2.18, 2.19, 3.7).

La définition de l'invariance d'un problème de décision est liée à des groupes de transformations dans les trois espaces participant à la définition d'un jeu statistique : les espaces D et Θ et l'espace des échantillons $\mathscr{X}^n$. Les transformations mesurables g de l'espace $\mathscr{X}^n$ forment un groupe G muni de l'opération de composition : si $g_1 \in G$ et $g_2 \in G$, alors $g_2 g_1$ est un élément de G tel que $x \to g_2(g_1 x)$. La transformation identique sera désignée par e.

La transformation g^{-1} réciproque de g se définit comme la transformation pour laquelle $g^{-1}g = e$. La mesurabilité de $g \in G$ exprime que gX et X seront des variables aléatoires dans $\mathscr{X}^n$.

La notion d'invariance d'une famille de distributions $\{\mathbf{P}_\theta\}$ qui a été définie dans les §§ 2.19 et 3.7 est étroitement liée au groupe G introduit. Elle exprime que pour tous $g \in G$ et $\theta \in \Theta$, on peut exhiber un élément $\theta_g \in \Theta$ tel que

$$\mathbf{P}_\theta(gX \in A) = \mathbf{P}_{\theta_g}(X \in A). \tag{2}$$

Les transformations $\bar{g} : \Theta \to \Theta$, définies par l'égalité $\bar{g}\theta = \theta_g$ sous la condition (A_0) forment un groupe $\overline{G}$ (cf. § 2.19).

En termes d'espérance mathématique, la condition (2) exprime que pour toute fonction intégrable φ, on a

$$\mathbf{E}_\theta\varphi(gX) = \mathbf{E}_{\bar{g}\theta}\varphi(X). \tag{3}$$

DÉFINITION 2. On dit que le problème de décision lié au jeu statistique $(\mathscr{D}, \Theta, w)$, $(X, \mathbf{P}_\theta)$ est *invariant par le groupe G* s'il en est de même et de la famille $\mathbf{P}_\theta$, et de la fonction de perte w au sens suivant : pour tous $\delta \in$ $\in D$ et $g \in G$, on peut exhiber un seul $\delta' \in D$ tel que

$$w(\delta, \theta) = w(\delta', \bar{g}\theta), \quad \forall\theta \in \Theta. \tag{4}$$

On désignera par $g'\delta$ la valeur δ' qui est définie de façon unique à l'aide de g.

LEMME 1. *Les transformations g' de l'espace D engendrées par le groupe G forment un groupe G'.*

DÉMONSTRATION. Nous montrerons que l'ensemble G' de toutes les transformations g' est stable pour la composition et de plus que $g_2'g_1' =$ $= (g_2g_1)'$.

En effet

$$w(\delta, \theta) = w(g_1'\delta, \bar{g}_1\theta) = w(g_2'g_1'\delta, \bar{g}_2\bar{g}_1\theta) = w\left((g_2g_1)'\delta, (\overline{g_2g_1})\theta\right).$$

Comme $(\overline{g_2g_1}) = \bar{g}_2\bar{g}_1$, on obtient pour raison d'unicité $(g_2g_1)' = g_2'g_1'$. ◄

Ainsi au groupe G des transformations g de $\mathscr{X}^n$ sont liés les groupes $\overline{G}$ et G' de transformations des espaces Θ et D. Le problème de décision est invariant par ces trois groupes de transformations. Il est donc naturel de choisir des décisions qui soient invariantes lorsqu'on passe d'un problème à un autre équivalent. L'adéquation de cette approche a été examinée en détail dans les §§ 2.18, 2.19 et 3.7.

DÉFINITION 3. On dit qu'une décision $\delta(X)$ d'un problème invariant est *invariante* si

$$\delta(gX) = g'\delta(X).$$

La *décision invariante randomisée* $\pi(X)$ se définit comme n'importe quelle distribution concentrée sur des décisions invariantes.

On trouvera des exemples d'application du principe d'invariance dans les §§ 2.18, 2.19, 3.7 où l'on a étudié les estimateurs équivariants et les tests invariants. Signalons un trait spécifique de ces deux cas particuliers.

Dans le *problème d'estimation*, le groupe de transformations G' n'a pas été introduit du tout. Dans ce cas les ensembles D et Θ sont confondus et dès le départ on a admis que $g'\delta = \bar{g}\delta$. C'est pourquoi les estimateurs équivariants ont été définis à l'aide de l'égalité $\theta^*(gX) = \bar{g}\theta^*(X)$.

En *théorie de test d'hypothèses*, la transformation g' a été supposée égale à la transformation identique $g' = e$, de sorte que le test invariant π a été défini par la relation $\pi(gX) = \pi(X)$.

Dans ce cas, pour que le problème de test des hypothèses $\{\theta \in \Theta_1\}$ et $\{\theta \in \Theta_2\}$ soit invariant, il faut admettre aussi (cf. (4)) que $g\Theta_i = \Theta_i$.

La différence entre ces deux approches justifie dans une certaine mesure l'utilisation de deux termes différents : l'équivariance (pour les estimateurs) et l'invariance (pour le test d'hypothèses) pour désigner des décisions invariantes. Nous avons envisagé plusieurs exemples de problèmes invariants dans les chapitres 2 et 3. Voici encore un autre.

EXEMPLE 1. Soit $X \in \Phi_{\alpha,\sigma^2}$. Prenons pour Θ le demi-plan $\{\theta = (\alpha, \sigma) : \sigma > 0\}$ et supposons que D est la droite réelle R et $w(\delta, \theta) = (\delta - \alpha)^2/\sigma^2$.

Considérons le groupe G de transformations $g_{a,b}X = a + bX = (a + bx_1, \ldots, a + bx_n)$, où $b \neq 0$. La variable aléatoire $g_{a,b}X$ de $\mathscr{X}^n$ peut visiblement être traitée comme un échantillon distribué suivant la loi $\Phi_{a+b\alpha,\, b^2\sigma^2}$. Donc, la famille Φ_{α,σ^2} est invariante par G si l'on pose $\bar{g}_{a,b}\theta = (a + b\alpha, |b|\sigma)$. La fonction de perte le sera aussi si l'on pose $g'_{a,b}\delta = a + b\delta$, puisque

$$w(g'_{a,b}\delta, \bar{g}_{a,b}\theta) = \frac{(a + b\delta - a - b\alpha)^2}{b^2\sigma^2} = w(\delta, \theta).$$

Nous avons donc affaire à un problème de décision invariant par G. Les décisions invariantes $\delta(X) : \mathscr{X}^n \to R$ doivent posséder la propriété

$$\delta(a + bX) = \delta(g_{a,b}X) = g'_{a,b}\delta(X) = a + b\delta(X). \tag{5}$$

Par ailleurs, on établit sans peine que ce problème de décision est invariant aussi par le groupe F des permutations f des coordonnées du vecteur X ; ceci étant, $\bar{f}$ et f' seront des transformations identiques. Si donc l'on exige que la fonction $\delta(X)$ soit une décision invariante par F aussi, il faudra alors que

$$\delta(fX) = \delta(X). \tag{6}$$

Signalons que la classe des fonctions vérifiant (5) et (6) est encore trop large : elle comprend par exemple toutes les formes linéaires

$$\delta(X) = \sum_{k=1}^{n} a_k x_{(k)}, \qquad \sum_{k=1}^{n} a_k = 1,$$

où $x_{(1)}, \ldots, x_{(n)}$ est l'échantillon ordonné associé à X. Si l'on fait intervenir le principe d'absence de biais, on obtient encore une condition sur les coefficients a_k :

$$\sum_{k=1}^{n} a_k \mathbf{E}_\theta(x_{(k)} - \alpha) = 0. \quad \blacktriangleleft$$

Les notions d'*orbite* (théorie de l'estimation) et d'*invariant* (théorie de test d'hypothèses), qui dans un certain sens sont voisines, jouent un rôle important dans la construction des décisions invariantes optimales. On rappelle qu'une orbite dans l'espace Θ est un ensemble $\{\bar{g}\theta_0, \bar{g} \in \overline{G}\}$, où θ_0 est un point arbitraire de Θ. En d'autres termes, θ_1 et θ_2 appartiennent à une même orbite s'il existe un $\bar{g} \in \overline{G}$ tel que $\theta_1 = \bar{g}\theta_2$.

On pourrait définir de façon analogue une orbite dans $\mathscr{X}^n$. Les invariants seraient alors par définition les statistiques constantes sur les orbites de $\mathscr{X}^n$.

La notion d'orbite garde sa signification dans le cas général aussi.

LEMME 2. *La fonction de risque d'un problème de décision invariant pour une décision invariante est constante sur une orbite :*

$$W(\delta(\cdot), \theta) = W(\delta(\cdot), \bar{g}\theta)$$

pour tous $\theta \in \Theta$, $\bar{g} \in \overline{G}$.

DÉMONSTRATION. L'invariance respectivement de la fonction de perte, de la décision et de la famille $\{\mathbf{P}_\theta\}$ (cf. (3), (4)) nous donne

$$W(\delta(\cdot), \theta) = \mathbf{E}_\theta w(\delta(X), \theta) = \mathbf{E}_\theta w(g'\delta(X), \bar{g}\theta) =$$
$$= \mathbf{E}_\theta w(\delta(gX), \bar{g}\theta) = \mathbf{E}_{\bar{g}\theta} w(\delta(X), \bar{g}\theta) = W(\delta(\cdot), \bar{g}\theta). \quad \blacktriangleleft$$

La constance, sur une orbite, du risque pour des décisions invariantes randomisées résulte de leur définition et du lemme 2.

Le lemme 2 nous dit alors que dans ce cas l'espace Θ tout entier sera une orbite (c'est-à-dire que $\Theta = \{\bar{g}\theta_0, \bar{g} \in \overline{G}\}$ pour un θ_0 quelconque ; ceci a lieu par exemple pour les translations) et la décision invariante devient niveleuse. Donc, le lemme 2 et les théorèmes 2.3 et 2.5 entraînent immédiatement la proposition suivante qui établit un lien important entre l'invariance et la minimaximalité.

THÉORÈME 2. *Supposons que l'espace Θ est une orbite et qu'il existe une distribution a priori* **Q** *pour laquelle la stratégie bayésienne $\pi_Q(X)$ est invariante. Alors $\pi_Q(X)$ est minimax.*

Le théorème 3.3 entraîne la généralisation suivante du théorème 2.

THÉORÈME 2A. *Supposons qu'il existe une distribution a priori* **Q** *concentrée sur une orbite de Θ_0 et telle que la stratégie bayésienne $\pi_Q(X)$ soit invariante. Si pour tous les θ*

$$W(\pi_Q(\cdot), \theta) \leqslant W(\pi_Q(\cdot), \theta_0), \quad \theta_0 \in \Theta_0,$$

alors $\pi_Q(X)$ est minimax.

Nous avons fait usage de ce test au § 3.9.

§ 6. Estimateurs asymptotiquement optimaux avec une fonction de perte arbitraire

De nombreux résultats du chapitre 2 sur les estimateurs asymptotiquement optimaux et du chapitre 3 sur les tests asymptotiquement optimaux peuvent être généralisés à des fonctions de perte de forme très générale.

Dans ce paragraphe on s'arrêtera sur des problèmes d'estimation et on admettra que $w(\delta, \theta) = w(\delta - \theta)$.

Faisons d'abord une remarque générale. Dans le chapitre 2 nous avons vu que dans le cas régulier ($X \in \mathbf{P}_\theta$, $\mathbf{P}_\theta$ satisfait les conditions (RR) ; cf. §§ 2.24, 2.28), tous les estimateurs $\theta^* = \delta(X)$ du paramètre θ étaient « concentrés » dans un $1/\sqrt{n}$-voisinage du point θ. Ainsi, par exemple, pour les estimateurs asymptotiquement normaux, $(\theta^* - \theta)\sqrt{n} \in \Phi_{0,\sigma^2(\theta)}$. Il s'ensuit que si l'on assujettit la fonction $w(t)$ à des conditions assez larges, le comportement asymptotique du risque $\mathbf{E}_\theta w(\theta^* - \theta)$ dépendra des propriétés de $w(t)$ au voisinage du point $t = 0$. Si $w(t)$ est bicontinûment dérivable en 0, $w'' > 0$, alors pour $t \to 0$

$$w(t) = \frac{w''(0)}{2} t^2 + o(t^2). \tag{1}$$

Ceci exprime que dans le domaine des valeurs de t (de l'ordre de $1/\sqrt{n}$) la fonction $w(t)$ se comportera comme la fonction de perte quadratique $w_0(t) = ct^2$, $c = w''(0)/2$, pour laquelle ont été établis les résultats du chapitre 2. Si de plus $w(t) < e^{\alpha|t|^2}$ pour un $\alpha > 0$ assez petit (cf. théorème 2.28.6), tous ces résultats restent en vigueur : leur extension à une fonction $w(t)$ de forme (1) est une tâche peu compliquée parfaitement accessible au lecteur.

Dans ce paragraphe, on se penchera sur une généralisation bien plus consistante. On admettra que la fonction de perte $w(\delta, \theta)$ dépend de n et se représente sous la forme

$$w(\delta, \theta) = w_n(\delta - \theta) = w(\sqrt{n}(\delta - \theta)), \tag{2}$$

où la fonction $w(t) \geqslant 0$ est définie dans l'espace R^k tout entier. Il est évident que $w(t)$ prendra, quel que soit t, des valeurs qui doivent être prises en considération.

On admettra que la fonction w de (2) satisfait les conditions suivantes :

1) *$w(t) \leqslant e^{c|t|}$ pour un certain $c > 0$.*

Cette forme de la condition 1) simplifie un peu les calculs. En effet, tous les résultats restent en vigueur si l'on exige que $w(t) \leqslant c_1 e^{\alpha|t|^2}$ pour $\alpha > 0$ assez petit.

La fonction

$$V_{\sigma^2}(s) = \int w(s - u) e^{-\frac{1}{2} u\sigma^2 u^T} du,$$

où σ^2 est une matrice des moments d'ordre deux définie positive, jouera un rôle très important dans la suite. Cette fonction peut être interprétée comme suit

$$V_{\sigma^2}(s) = \frac{(2\pi)^{k/2}}{\sqrt{|\sigma^2|}} \mathsf{E} w(s - \xi), \quad \xi \in \Phi_{0,\sigma^{-2}}.$$

C'est une fonction analytique de s et de σ^2, puisque

$$V_{\sigma^2}(s) = \int w(v) e^{-\frac{1}{2}(s - v)\sigma^2(s - v)^T} dv.$$

2) *La fonction $V_{\sigma^2}(s)$ atteint son minimum par rapport à s en un seul point que l'on désignera par b_w.*

3) $b_w = 0$.

4) *La fonction $w(t)$ est continue.*

La condition 2) sera visiblement satisfaite si $w(s) \neq$ const est une fonction convexe vers le bas. Il est évident que $V_{\sigma^2}(s)$ sera aussi convexe et ne comportera pas de portions « linéaires » (c'est-à-dire que la matrice des dérivées secondes sera partout définie positive).

La condition 3) sera remplie si

$$V'_{\sigma^2}(0) = - \int u w(u) e^{-\frac{1}{2} u\sigma^2 u^T} du = 0,$$

ce qui a toujours lieu pour les fonctions symétriques $w(u) = w(-u)$.

On aurait pu appeler la valeur b_w *biais* de la fonction de perte w. Cette valeur b_w vérifie l'équation $V'_{\sigma^2}(b_w) = 0$. La condition 3) de nullité de b_w n'est pas essentielle et son seul objectif est de simplifier l'exposé. Le lecteur

pourra traiter sans peine le cas $b_w \neq 0$. Les changements qui interviendront dans les énoncés des théorèmes seront illustrés dans la remarque 2 qui suit le théorème 1.

Rappelons ce que deviendront les définitions des stratégies optimales des §§ 2 et 3. Un estimateur θ^*_Q sera *bayésien* par rapport à une distribution *a priori* **Q** de densité q par rapport à la mesure de Lebesgue (et à la fonction de perte w_n) si

$$\int W(\theta^*_Q, t)q(t)dt = \min_{\theta^*}\int W(\theta^*, t)q(t)dt, \tag{3}$$

où $W(\theta^*, t) = \mathbf{E}_t w_n(\theta^* - t)$. L'intégrale du second membre peut être mise sous la forme de l'espérance mathématique $\mathbf{E}w_n(\theta^* - \theta)$ où la moyenne est prise par rapport à une distribution de densité $f_t(x)q(t)$.

Un estimateur $\bar{\theta}^*$ est *minimax* si pour tout autre estimateur θ^*

$$\sup_t W(\bar{\theta}^*, t) \leqslant \sup_t W(\theta^*, t).$$

Ceci nous suggère tout naturellement les définitions suivantes qui sont calquées sur celles du § 2.11.

DÉFINITION 1. On dit qu'un estimateur θ^* est *asymptotiquement bayésien* si

$$\lim_{n\to\infty}\text{.sup}[\mathbf{E}w_n(\theta^* - \theta) - \mathbf{E}w_n(\theta^*_Q - \theta)] \leqslant 0, \tag{4}$$

où θ^*_Q est un estimateur bayésien.

DÉFINITION 2. On dit qu'un estimateur θ^*_1 est *asymptotiquement minimax* si pour tout autre estimateur θ^*

$$\lim_{n\to\infty}\text{.sup}\left[\sup_{t\in\Theta_0} W(\theta^*_1, t) - \sup_{t\in\Theta_0} W(\theta^*, t)\right] \leqslant 0, \tag{5}$$

où $\Theta_0 \subset \Theta$ est un sous-ensemble fermé quelconque.

Pour étudier les estimateurs asymptotiquement optimaux, on s'appuyera dans ce paragraphe sur les seules notions introduites par les définitions 1 et 2. Ceci tranchera avec le chapitre 2 qui faisait intervenir aussi les estimateurs asymptotiquement efficaces. L'absence de ces derniers s'explique par le fait que pour les fonctions de perte w arbitraires, nous ne disposons pas d'inégalités de Rao-Cramer pour $\inf_{\theta^*\in K_0} W(\theta^*, \theta)$ (K_0 est la classe des estimateurs sans biais) qui nous permettent de juger de la qualité de θ^* d'après la valeur de $W(\theta^*, \theta)$ et de déterminer, en particulier, les estimateurs efficaces (et asymptotiquement efficaces), c'est-à-dire les estimateurs uniformément les meilleurs dans la classe K_0.

Les propositions suivantes expriment que l'estimateur du maximum de vraisemblance est, comme dans les conditions du chapitre 2, asymptotiquement bayésien et asymptotiquement minimax. Nous déterminerons par ail-

leurs la borne inférieure *asymptotique* de la fonction de risque pour une fonction de perte quelconque w (l'inégalité de Rao-Cramer nous fournit la borne inférieure *exacte*). Dans les trois théorèmes ultérieurs, on admet que les conditions (RR) sont remplies.

THÉORÈME 1. *Supposons que* $X \in \mathbf{P}_\theta$, $\hat{\theta}^*$ *est un estimateur du maximum de vraisemblance et* $\theta^*_{\mathbf{Q}}$ *l'estimateur bayésien associé à une fonction de perte* w (cf. (2)) *satisfaisant les conditions* 1), 2) *et* 3) *et à une distribution a priori* $\mathbf{Q}$ *de densité bornée* q *par rapport à la mesure de Lebesgue. Alors*

$$|\theta^*_{\mathbf{Q}} - \hat{\theta}^*| \sqrt{n} \underset{P_\theta}{\to} 0, \tag{6}$$

$$(\theta^*_{\mathbf{Q}} - \theta) \sqrt{n} \Subset \Phi_{0, I^{-1}(\theta)} \tag{7}$$

uniformément en $\theta \in \Theta_0$, $\Theta_0 \subset \Theta$ *étant un sous-ensemble fermé sur lequel* $q(\theta) > q_0 > 0$ *est continue.*

Si, de plus, la fonction w *vérifie la condition* (4), *alors*

$$\mathsf{E} w_n(\theta^*_{\mathbf{Q}} - \theta) = \mathsf{E} w(\sqrt{n}(\theta^*_{\mathbf{Q}} - \theta)) \to \mathsf{E} w(\eta_\theta) = \mathsf{E} \frac{\sqrt{I(\theta)}}{(2\pi)^{k/2}} V_{I(\theta)}(0), \tag{8}$$

où $\eta_\theta \in \Phi_{0, I^{-1}(\theta)}$, $\theta \in \mathbf{Q}$; E *désigne comme précédemment l'espérance mathématique par rapport à la densité* $f_t(x) q(t)$ ($X \in \mathbf{P}_\theta$, $\theta \in \mathbf{Q}$).

REMARQUE 1. Conjointement à la convergence (6), on peut établir la convergence presque sûre pour la mesure $\mathbf{P}_\theta$.

REMARQUE 2. Si w est telle que le biais $b_w \neq 0$, le théorème 1 reste entièrement en vigueur, pourvu que l'on remplace $\theta^*_{\mathbf{Q}}$ par $\theta^*_{\mathbf{Q}} - b_w/\sqrt{n}$ dans (6), (7) et (8). Donc, b_w s'interprète comme le *biais asymptotique de* $(\theta^*_{\mathbf{Q}} - \theta)\sqrt{n}$.

THÉORÈME 2. *Supposons que la fonction* w *satisfait les conditions* 1) *à* 4). *Alors, pour tout estimateur* θ^*

$$\liminf_{n \to \infty} \sup_{t \in \Theta_0} \mathsf{E}_t w_n(\theta^* - t) \geqslant \sup_{t \in \Theta_0} \mathsf{E} w(\eta_t), \tag{9}$$

$$\eta_t \in \Phi_{0, I^{-1}(t)}.$$

Tout estimateur θ^* *tel que*

$$\mathsf{E}_t w_n(\theta^* - t) \to \mathsf{E} w(\eta_t) \tag{10}$$

uniformément en à t, *est asymptotiquement minimax.*

THÉORÈME 3. *Supposons que* $X \in \mathbf{P}_\theta$ *et que la fonction* w *satisfait les conditions* 1) *à* 4). *L'estimateur du maximum de vraisemblance* $\hat{\theta}^*$ *est alors*

asymptotiquement minimax et asymptotiquement bayésien pour toute distribution a priori **Q** *de densité q continue, strictement positive au point* θ.

Ces propositions sont identiques aux propositions correspondantes du chapitre 2. Elles rendent vraisemblable l'hypothèse que pour toute fonction de perte w vérifiant les conditions 1) à 4) l'estimateur du maximum de vraisemblance est uniformément et asymptotiquement le meilleur dans la classe des estimateurs asymptotiquement sans biais (comparer avec les §§ 2.25 et 2.28).

DÉMONSTRATION du théorème 1. L'estimateur bayésien se définit en vertu du principe de Bayes comme un estimateur dont la valeur θ_Q^* est telle que

$$\int w_n(\theta_Q^* - t)q(t|x)dt = \min_{u\in\Theta}\int w_n(u - t)q(t|X)dt =$$
$$= \min_{u\in\Theta}\int w(\sqrt{n}(u - \theta) - \sqrt{n}(t - \theta)\frac{q(t)f_t(X)}{\int q(v)f_v(X)dv}dt.$$

Ceci exprime que pour $(\theta_Q^* - \theta)\sqrt{n} \equiv u_Q^*$ on peut prendre n'importe quelle valeur s pour laquelle est atteint $\min\limits_s U(s)$,

$$U(s) = \int w(s - v)q\left(\theta + \frac{v}{\sqrt{n}}\right)Z\left(\frac{v}{\sqrt{n}}\right)dv, \tag{11}$$

où comme précédemment $Z(t) = \dfrac{f_{\theta+t}(X)}{f_\theta(X)}$.

Nous aurons besoin de propositions relatives au comportement asymptotique de $U(s)$. Dans les §§ 2.28 et 2.29, nous avons établi (théorème 2.28.5) que si les conditions (RR) étaient réunies, alors

$$U(u^*) = e^{Y(u^*)}q(\hat\theta^*)(V_{I(\theta)}(0) + \varepsilon_n(X, \theta)), \tag{12}$$

où $\varepsilon_n(X, \theta) \underset{P_\theta}{\to} 0$ uniformément en θ (nous avons remplacé ici $\dfrac{(2\pi)^{k/2}}{\sqrt{I(\theta)}}\times$ $\times \mathbf{E}w(\xi)$ par $V_{I(\theta)}$, et $q(\theta)$ par $q(\hat\theta^*)$).

Remarquons maintenant que

$$\mathbf{P}(\sqrt{n}|\theta_Q^* - \hat\theta^*| \geqslant \varepsilon) = \mathbf{P}(|u_Q^* - u^*| \geqslant \varepsilon) \leqslant$$
$$\leqslant \mathbf{P}\Big(\min_{|s-u^*|\geqslant\varepsilon} U(s) \leqslant U(u^*)\Big). \tag{13}$$

Vu que nous connaissons la représentation asymptotique de $U(u^*)$ il nous faut estimer la valeur $U(s)$. Des théorèmes 2.28.4 et 2.29.3 il s'ensuit

que pour toute suite arbitraire $\delta_n \to 0$, pour $|v| < \delta_n\sqrt{n}$

$$\ln Z\left(\frac{v}{\sqrt{n}}\right) = Y(u^*) - \frac{1}{2}(v - u^*)I(\theta)(v - u^*)^T(1 + \epsilon_n(X, \theta, u)),$$

$|\epsilon_n(X, \theta, u)| \leqslant \epsilon_n^{(1)}(X, \theta) \underset{P_\theta}{\to} 0$ uniformément en θ. Mais

$$U(s) \geqslant U_n(s) \equiv \int\limits_{|v - u^*| < \delta_n\sqrt{n}} w(s - v)q\left(\theta + \frac{v}{\sqrt{n}}\right)Z\left(\frac{v}{\sqrt{n}}\right)dv.$$

Considérons l'ensemble

$$A_n = \left\{\epsilon_n^{(1)}(X, \theta) < \varrho, \inf_{|v - u^*| < \delta_n\sqrt{n}} q\left(\theta + \frac{v}{\sqrt{n}}\right) > q(\hat{\theta}^*)(1 - \varrho)\right\},$$
$$\varrho > 0,$$

pour lequel de toute évidence

$$\mathbf{P}_\theta(A_n) \to 1. \tag{14}$$

Sur cet ensemble, on a uniformément en θ

$$U_n(s) \geqslant (1 - \varrho)q(\hat{\theta}^*)e^{Y(u^*)} \times$$
$$\times \int\limits_{|v - u^*| < \delta_n\sqrt{n}} w(s - v)\exp\left\{-\frac{1}{2}(v - u^*)I(\theta)(v - u^*)^T(1 + \varrho)\right\}dv =$$
$$= (1 - \varrho)q(\hat{\theta}^*)e^{Y(u^*)}[V_{I(\theta)(1+\varrho)}(s - u^*) - r_n(s)], \tag{15}$$

où en vertu de la condition 1)

$$r_n(s) = \int\limits_{|v - u^*| > \delta_n\sqrt{n}} w(s - v)\exp\left\{-\frac{1}{2}(v - u^*)I(\theta)(v - u^*)^T \times\right.$$
$$\left.\times (1 + \varrho)\right\}dv \leqslant e^{c\sqrt{nd}} \frac{(2\pi)^{k/2}}{\sqrt{|I(\theta)|c}}\mathbf{P}\left(|\eta| > \delta_n\sqrt{n}\right),$$
$$\eta \in \Phi_{0, I(\theta)(1+\varrho)},$$

d étant le diamètre du domaine Θ. Comme pour le lemme 2.23.1 on s'assure sans peine que

$$\mathbf{P}\left(|\eta| > \delta_n\sqrt{n}\right) \leqslant e^{-\alpha n\delta_n^2}, \ \alpha > 0.$$

En choisissant $\delta_n = n^{-1/9}$, on trouve que pour tous les s et les n assez grands

$$r_n(s) \leqslant e^{-n^{3/4}}. \tag{16}$$

Appliquons maintenant les conditions 2) et 3) en vertu desquelles

$$\min_{|s-u^*|>\epsilon} V_{I(\theta)}(s - u^*) \geqslant V_{I(\theta)}(0) + 4\tau,\ \tau = \tau(\epsilon) > 0.$$

D'après les propriétés analytiques de $V_{\sigma^2}(s)$, on obtient pour les ϱ assez petits

$$\min_{|s-u^*|>\epsilon} V_{I(\theta)(1+\varrho)}(s - u^*) \geqslant V_{I(\theta)}(0) + 3\tau,$$

et d'après (15) et (16) pour $X \in A_n$ et pout n assez grand

$$\min_{|s-u^*|>\epsilon} U_n(s) \geqslant (1 - \varrho)q(\hat{\theta}^*)e^{Y(u^*)}[Y_{I(\theta)}(0) + 2\tau].$$

En se servant de (12) et (13), on trouve en définitive

$$\mathbf{P}_\theta\left(\sqrt{n}|\theta^*_Q - \hat{\theta}^*| \geqslant \epsilon\right) \leqslant \mathbf{P}_\theta(\min_{|s-u^*|>\epsilon} U_n(s) \leqslant U(u^*)) \leqslant$$

$$\leqslant \mathbf{P}_\theta\left(X \notin A_n\right) + \mathbf{P}_\theta((1 - \varrho)[V_{I(\theta)}(0) + 2\tau] \leqslant$$

$$\leqslant V_{I(\theta)}(0) + \epsilon_n(X, \theta)).$$

En choisissant ϱ suffisamment petit pour que $(1 - \varrho)2\tau - \varrho V_{I(\theta)}(0) \geqslant \geqslant \tau$, on obtient

$$\mathbf{P}\left(\sqrt{n}|\theta^*_Q - \hat{\theta}^*| \geqslant \epsilon\right) \leqslant \mathbf{P}_\theta(X \notin A_n) + \mathbf{P}_\theta(\epsilon_n(X, \theta) > \tau) \to 0$$

lorsque $n \to \infty$. Ce qui prouve la proposition (6) en vertu de (12) et (14).

La relation (7) découle de (6) et des théorèmes du § 2.29. Prouvons maintenant la relation (8). En vertu de (7) et de la propriété 4) il vient

$$w\left(\sqrt{n}(\theta^*_Q - \theta)\right) \Rightarrow w(\eta_\theta),\ \eta_\theta \in \Phi_{0,I^{-1}(\theta)}.$$

Le lemme de Fatou nous donne

$$\liminf_{n\to\infty} \mathbf{E}_t w\left(\sqrt{n}(\theta^*_Q - t)\right) \geqslant \mathbf{E}w(\eta_t),$$

$$\liminf_{n\to\infty} \mathbf{E}w\left(\sqrt{n}(\theta^*_Q - \theta)\right) \geqslant \int q(t)\mathbf{E}w(\eta_t)dt = \mathbf{E}w(\eta_\theta).$$

Par ailleurs, par définition de θ^*_Q on a

$$\mathbf{E}w\left(\sqrt{n}(\theta^*_Q - (\theta)\right) \leqslant \mathbf{E}w\left(\sqrt{n}(\hat{\theta}^* - \theta)\right) \to \mathbf{E}w(\eta_\theta).$$

La dernière relation résulte de la convergence uniforme de $\mathbf{E}_t w(\sqrt{n}(\hat{\theta}^* - - t)) \to \mathbf{E}w(\eta_t)$ qui a été établie dans le § 2.29. ◄

DÉMONSTRATION du théorème 2. Prenons une distribution $\mathbf{Q}$ concentrée sur Θ_0, de densité $q(t) > 0$ bornée pour $t \in \Theta_0$ et supposons que θ^*_Q

est l'estimateur bayésien associé à $\mathbf{Q}$. Alors pour tout estimateur θ^*

$$\sup_{t\in\Theta_0} \mathbf{E}_\theta w_n(\theta^* - t) \geqslant \int_{\Theta_0} \mathbf{E}_t w_n(\theta^* - t)q(t)dt \geqslant$$

$$\geqslant \int_{\Theta_0} \mathbf{E}_t w_n(\theta^*_{\mathbf{Q}} - t)q(t)dt = \mathbf{E} w_n(\theta^*_{\mathbf{Q}} - \theta).$$

Le lemme de Fatou nous donne, eu égard à (8),

$$\lim_{n\to\infty}\text{.inf.}\sup_{t\in\Theta_0}\text{.}\mathbf{E}_t w_n(\theta^* - t) \geqslant \lim_{n\to\infty}\text{.inf}\,\mathbf{E} w_n(\theta^*_{\mathbf{Q}} - \theta) \geqslant \mathbf{E} w(\eta_\theta) =$$

$$= \int_{\Theta_0} \mathbf{E} w(\eta_t)q(t)dt.$$

La fonction $\mathbf{E} w(\eta_t) = \dfrac{\sqrt{I(t)}}{(2\pi)^{k/2}} V_{I(t)}(0)$ étant continue, par rapport à t,

l'intégrale

$$\int_{\Theta_0} \sqrt{I(t)}\, V_{I(\theta)}(0)q(t)dt$$

peut être rendue aussi proche que l'on veut de $\sup_{t\in\Theta_0}\sqrt{I(t)}\, V_{I(t)}(0) =$ $= \sup_{t\in\Theta_0} \mathbf{E} w(\eta_t)$ par un choix convenable de $q(t)$. Ce qui prouve (9).

Supposons maintenant qu'un estimateur θ_1^* possède la propriété (10) et soit θ^* un autre estimateur quelconque. En vertu de (9) et de la convergence uniforme de (10), il vient

$$\lim_{n\to\infty}\text{.sup}\left[\sup_{t\in\Theta_0}\mathbf{E}_t w_n(\theta_1^* - t) - \sup_{t\in\Theta_0}\mathbf{E}_t w_n(\theta^* - t)\right] \leqslant$$

$$\leqslant \sup_{t\in\Theta_0}\text{.}\lim_{n\to\infty} \mathbf{E}_t w_n(\theta_1^* - t) - \sup_{t\in\Theta_0}\mathbf{E} w(\eta_t) = 0.$$

On a ainsi prouvé l'inégalité (5) et avec elle le théorème 2.

DÉMONSTRATION du théorème 3. Que $\hat{\theta}^*$ soit asymptotiquement minimax résulte de ce qu'il est justiciable de (10) en vertu du théorème 2.29.4.

Qu'il soit asymptotiquement bayésien résulte de ce que la relation (4) est valable pour $\theta^* = \hat{\theta}^*$, car $\hat{\theta}^*$ est justiciable de la convergence uniforme (10) et par suite

$$\lim_{n\to\infty} \mathbf{E} w_n(\hat{\theta}^* - \theta) = \lim_{n\to\infty}\int \mathbf{E}_t w_n(\hat{\theta}^* - \theta)q(t)dt =$$

$$= \mathbf{E} w(\eta_\theta) = \lim_{n\to\infty} \mathbf{E} w_n(\theta^*_{\mathbf{Q}} - \theta).$$

La dernière égalité découle de (8). Ce que nous voulions.

On peut renforcer le théorème 1 en exigeant accessoirement que la fonction $w(t)$ soit à croissance assez rapide. Plus exactement, posons $w_N = \min_{|t|>N} w(t)$ et $W_M = \max_{|t|\le M} w(t)$ et considérons la condition

5) *Il existe* $\gamma < 1$ *tel que* $w_N > 2W_{\gamma N}$ *pour tous les* N *assez grands.*

Si $w(t)$ croît comme une fonction puissance ou exponentielle lorsque $|t| \to \infty$, la condition 5) est satisfaite.

THÉORÈME 4. *Si les conditions* 1) *et* 5) *sont satisfaites,* $q(t) > q_0 > 0$ *sur un ensemble fermé* Θ_0 *et* $q(t) \leqslant q_m < \infty$, *alors pour des* $c < \infty$ *et* $\alpha > 0$ *indépendants de* t, *on a*

$$\mathbf{P}_t\left(\sqrt{n}\left(\theta_{\mathfrak{Q}}^* - t\right) > N\right) \leqslant ce^{-\alpha N^2},\ t \in \Theta_0.$$

Ceci et le théorème 1 entraînent que pour toute fonction $v(t)$ continue telle que $|v(t)| \leqslant e^{-\alpha N^2/2}$, on a

$$\mathbf{E}_t v\left(\sqrt{n}\left(\theta_{\mathfrak{Q}}^* - t\right)\right) \to \mathbf{E}v(\eta_t),\ t \in \Theta_0.$$

Posons

$$u(r) = \int_{|v|\geqslant r} w(-v)q\left(\theta + \frac{v}{\sqrt{n}}\right)Z\left(\frac{v}{\sqrt{n}}\right)dv$$

(ceci est la partie de l'intégrale $U(0)$, étendue au domaine $|v| \geqslant r$). Pour prouver le théorème 4, nous aurons besoin du

LEMME 1. *Si* $w(t)$ *satisfait la condition* 1), $q_m = \max_u q(u) < \infty$, *alors pour certains* $\beta > 0$, $a < \infty$, *indépendants de* θ *et pour tous les* $0 < \delta < 1$, *on a*

$$\mathbf{P}_\theta(u(r) > \delta) \leqslant \frac{a}{\delta}e^{-\beta r^2}.$$

Cette inégalité est valable pour $w(t) \equiv 1$.

DÉMONSTRATION. On a

$$\mathbf{P}_\theta(u(r) > \delta) \leqslant \mathbf{P}_\theta\left(\sup_{|v|\geqslant r} Z\left(\frac{v}{\sqrt{n}}\right) > 1\right) + \mathbf{P}_\theta\left(u(r) > \delta,\ \sup_{|v|\geqslant r} Z\left(\frac{v}{\sqrt{n}}\right) \leqslant 1\right).$$

Dans le théorème 2.23.2 on a vu que le premier terme est majoré par $c_1e^{-r^2\beta}$, $\beta > 0$. Le second terme est majoré par

$$\mathbf{P}_\theta\left(\int_{|v|\geqslant r} w(-v)q\left(\theta + \frac{v}{\sqrt{n}}\right)Z^{1/2}\left(\frac{v}{\sqrt{n}}\right)dv > \delta\right). \tag{17}$$

Vu que

$$\mathbf{E}_\theta Z^{1/2}\left(\frac{v}{\sqrt{n}}\right) \leqslant e^{-2|v|^2\beta},\ \beta > 0,$$

en vertu du théorème 2.23.1, l'espérance mathématique de l'intégrale de (17) est majorée par (cf. lemme 2.23.1)

$$q_m \int_{|v|\geq r} e^{c|v|} e^{-a|v|^{2\beta}} dv \leq c_2 e^{-r^{2\beta}}.$$

Donc, en vertu de l'inégalité de Tchébychev, la probabilité (17) est au plus égale à $c_2 e^{-r^{2\beta}}/\delta$. ◄

Désignons par $u_1(r)$ la valeur de l'intégrale $u(r)$ pour $w(t) \equiv 1$:

$$u_1(r) = \int_{|v|\geq r} q\left(\theta + \frac{v}{\sqrt{n}}\right) Z\left(\frac{v}{\sqrt{n}}\right) dv.$$

LEMME 2. *Si $q(\theta) > 0$ sur un ensemble fermé Θ_0, alors*

$$\mathbf{P}_\theta(u_1(0) < \varepsilon) \leq b\varepsilon^2, \ \theta \in \Theta_0,$$

pour un certain $b < \infty$ indépendant de θ, $\varepsilon > 0$ quelconque et tous les n assez grands.

DÉMONSTRATION. Pour tous les n assez grands

$$u_1(0) \geq \int_{|v|\leq 1} q\left(\theta + \frac{v}{\sqrt{n}}\right) Z\left(\frac{v}{\sqrt{n}}\right) dv \geq$$

$$\geq q_0 \int_{|v|\leq 1} \exp\left\{L\left(X, \theta + \frac{v}{\sqrt{n}}\right) - L(X, \theta)\right\} dv =$$

$$= q_0 \int_{|v|\leq 1} \exp\left\{(v, \zeta_n) + \frac{1}{2} v\gamma_n v^T\right\} dv,$$

où

$$q_0 = \min_{t\in\Theta_0} q(t) > 0, \ \zeta_n = \frac{1}{\sqrt{n}} L'(X, t), \ \gamma_n = \frac{1}{n} \left\| L''_{ij}(X, \tilde{\theta}) \right\|,$$

$\tilde{\theta} = \theta + \varrho v n^{-1/2}$, $|\varrho| \leq 1$. (L' est le vecteur des dérivées du logarithme de la fonction de vraisemblance, L''_{ij} sont les dérivées partielles du second ordre.) Comme $|(v, \zeta_n)| \leq |v|\,|\zeta_n|$ et que, en vertu des conditions (*RR*),

$$\left|v\gamma_n v^T\right| \leq \frac{1}{n} \sum_{i=1}^{n} l(x_i) \sum_{i,j=1}^{k} \left|v_i v_j\right| \leq \frac{k|v|^2}{n} L_n,$$

où $L_n = \sum_{i=1}^{n} l(x_i)$, on a sur l'ensemble $A = \{|\zeta_n| \leq 1/\epsilon, \ L_n \leq n/\epsilon^2 k\}$

$$u_1(0) \geq q_0 \int_{|v|\leq 1} \exp\left\{-\frac{|v|}{\epsilon} - \frac{|v|^2}{2\epsilon^2}\right\} dv \geq q_0 \int_{|v|\leq \epsilon^{-1}} \exp\left\{-|s| - \frac{|s|^2}{2}\right\} ds \geq c_1\epsilon.$$

Ceci exprime que $\{u_1(0) < c_1\epsilon\} \subset \overline{A}$. Puisque

$$\mathbf{P}_\theta(\overline{A}) \leq \mathbf{P}_\theta(|\zeta_n| \geq \epsilon^{-1}) + \mathbf{P}_\theta\left(L_n > \frac{n}{\epsilon^2 k}\right) \leq \epsilon^2 \mathbf{E}_\theta|\zeta_n|^2 + \frac{\epsilon^2 k}{n} \mathbf{E}_\theta L_n,$$

$$\mathbf{E}_\theta|\zeta_n|^2 = \sum_{l=1}^{k} I_{ll}(\theta),\ \mathbf{E}_\theta L_n = n\mathbf{E}_\theta l(x_1),$$

il vient

$$\mathbf{P}_\theta(\overline{A}) \leqslant c_2\varepsilon^2. \quad ◄$$

DÉMONSTRATION du théorème 4. Désignons par M_n l'ensemble des points s en lesquels est réalisé $\min U(s)$ (c'est-à-dire l'ensemble des points $(\theta^*_Q - \theta)\sqrt{n}$; cf. (34)) *). Alors

$$\{M_n \subset D\} = \left\{\min_{s\in D} U(s) < \min_{s\notin D} U(s)\right\}. \tag{18}$$

Donc

$$\{\sqrt{n}|\theta^*_Q - \theta| > 2N\} = \left\{\min_{|s|>2N} U(s) < \min_{|s|\leqslant 2N} U(s)\right\} \subset \left\{\min_{|s|\geqslant 2N} U(s) < U(0)\right\}.$$

Ici

$$\min_{|s|>\geqslant 2N} U(s) \geqslant w_n \int_{|u|<N} q\left(\theta + \frac{u}{\sqrt{n}}\right) Z\left(\frac{u}{\sqrt{n}}\right) du = w_N\left(u_1(0) - u_1(N)\right),$$

$$w_N \equiv \min_{\substack{|s|>2N\\|u|<N}} w(s-u) = \min_{|t|>N} w(t).$$

D'autre part

$$U(0) = \int w(-u) q\left(\theta + \frac{u}{\sqrt{n}}\right) Z\left(\frac{u}{\sqrt{n}}\right) du \leqslant (u_1(0) - u_1(M))W_M + u(M),$$

où $W_M = \max_{|t|<M} w(t)$.

De là on déduit

$$\left\{\sqrt{n}\ \middle|\ \theta^*_Q - \theta\ \middle|\ > 2N\right\} \subset \left\{w_N\left(u_1(0) - u_1(N)\right) < W_M\left(u_1(0) - u_1(M)\right) + u(M)\right\} \subset$$

$$\subset \left\{\left(\frac{w_N}{W_M} - 1\right) u_1(0) < \frac{u(M)}{W_M} + \frac{u_1(N) w_N}{W_M} + u_1(M)\right\}.$$

En vertu de la condition 5) choisissons $M = \gamma N$, $\gamma < 1$, de telle sorte que $w_N > 2W_M$ pour tous les N assez grands. Utilisons par ailleurs les inégalités $W_M > 2$ (pour les M assez grands), $w_N < w(N) < e^{cN}$. Il est alors évident que

$$\{\sqrt{n}\,|\,\theta^*_Q - \theta\,|\, > 2N\} \subset \left\{u_1(0) < u(\gamma N) + u_1(N)e^{cN}\right\}. \tag{19}$$

Le lemme 1 nous donne

$$\mathbf{P}_\theta\left(u(\gamma N) > \frac{1}{2} e^{-\alpha N^2}\right) \leqslant 2ae^{-\beta N^2\gamma^2 + \alpha N^2},$$

$$\mathbf{P}_\theta\left(u_1(N) > \frac{1}{2} e^{-cN-\alpha N^2}\right) \leqslant 2ae^{-\beta N^2 + \alpha N^2 + cN}.$$

*) A la place de M_n on aurait par exemple pu considérer le point de plus petite norme en lequel est réalisé $\min U(s)$.

En prenant $\alpha < \frac{1}{2}\beta\gamma^2$, on trouve que pour les grands N, la relation (19) entraîne

$$\mathbf{P}_\theta\left(\sqrt{n}\ \left|\ \theta^*_Q - \theta\ \right| > 2N\right) \leqslant 4ae^{-\alpha N^2} + \mathbf{P}_\theta\left(u_1(0) < e^{-\alpha N^2}\right).$$

Reste à appliquer le lemme 2 en vertu duquel

$$\mathbf{P}_\theta\left(u_1(0) < e^{-\alpha N^2}\right) \leqslant be^{-2\alpha N^2}. \blacktriangleleft$$

§ 7. Tests optimaux avec une fonction de perte arbitraire. Test du rapport de vraisemblance traité comme une décision asymptotiquement bayésienne

1. Optimalité des tests statistiques avec une fonction de perte arbitraire. Nous avons vu dans les deux paragraphes précédents que de nombreux résultats fondamentaux de la théorie de l'estimation s'étendaient qualitativement à des problèmes plus généraux de décision statistique avec des pertes $w(\delta, \theta)$, $\delta \in D \subset R^k$, $\theta \in \Theta \subset R^k$ non quadratiques.

On retrouve le même tableau en théorie des tests d'hypothèses. Nous avons vu au § 4 que les décisions optimales pour les jeux à ensembles D et Θ finis et à fonction de perte arbitraire étaient de la même forme que les tests optimaux d'un nombre fini d'hypothèses simples envisagés dans le § 3.1. Les résultats des §§ 3.5, 3.6, 3.7, 3.9, 3.11, 3.13, 3.14 et 3.15 sont valables aussi pour l'essentiel. En particulier, les théorèmes relatifs aux tests uniformément les plus puissants des §§ 3.5, 3.6 et 3.7 se transformeront en propositions pour les stratégies uniformément les meilleures dans les jeux statistiques correspondants ($\Theta \subset R^k$, $D = \{\delta_1, \delta_2\}$) dans lesquels toutefois la fonction de perte $w(\delta_i, \theta) = w_i(\theta)$, $w_i(\theta) = 0$ pour $\theta \in \Theta_i$, $i = 1, 2$, ne sera pas nécessairement statistique ($w_i(\theta) = 1$ pour $\theta \notin \Theta_i$) mais satisfera seulement certaines conditions assez générales (par exemple sera monotone croissante lorsque θ s'éloignera de Θ_i). Le rôle des classes K_ε dans lesquelles nous avons cherché les tests uniformément les plus puissants sera tenu dorénavant par les classes de décisions $\pi(X)$ dont la valeur maximale ε des « pertes de première espèce » est fixée :

$$\varepsilon = \sup_{\theta\in\Theta_1} W(\pi(\cdot), \theta) = \sup_{\theta\in\Theta_1} w_2(\theta)\mathbf{E}_\theta\pi(X, \delta_2). \qquad (1)$$

On minimisera les « pertes de deuxième espèce » :

$$W(\pi(\cdot), \theta) = w_1(\theta)\mathbf{E}_\theta\pi(X, \delta_1), \ \theta \in \Theta_2. \qquad (2)$$

Ici $\pi(X, \delta_i)$ désigne la probabilité d'accepter la décision δ_i[2] par le test π. Pour simplifier les notations, on posera, suivant le chapitre 3, $\pi(X, \delta_2) = \pi(X)$, de sorte que $\pi(X, \delta_1) = 1 - \pi(X)$. La désignation du test et du nombre $\pi(X, \delta_2)$ par le même symbole $\pi(X)$ est commode, et comme nous l'avons vu dans les chapitres précédents, ne prête pas à équivoque.

Dans (1) et (2) on cherche les extrémums d'expressions qui ne diffèrent des expressions homologues pour fonctions de perte statistiques que par des facteurs multiplicatifs indépendants de $\pi(X)$. Si ces facteurs sont monotones, l'exposé des §§ 3.5 à 3.7, 3.9 et 3.11 ne subit pas de changements notables lorsqu'on passe au problème défini par (1) et (2).

Les résultats à caractère asymptotique des §§ 3.13 à 3.15 seront peu modifiés aussi. Dans ce paragraphe on se penchera plus en détail sur la généralisation des résultats du § 3.13 au cas d'une fonction de perte arbitraire et l'on verra que cette généralisation ne nécessite aucun effort supplémentaire.

2. Test du rapport de vraisemblance traité comme un test asymptotiquement bayésien. Considérons un jeu statistique $(\mathscr{D}, \Theta, W)$ dans lequel l'ensemble Θ est un compact convexe de R^k ayant la puissance du continu et l'ensemble D des stratégies est un doubleton : $D = \{\delta_1, \delta_2\}$. La fonction de perte $w(\delta, \theta)$ est de la forme

$$w(\delta_1, \theta) = \begin{cases} 0, & \theta = \theta_1, \\ w_1(\theta), & \theta \neq \theta_1, \end{cases}$$

$$w(\delta_2, \theta) = \begin{cases} w_2, & \theta = \theta_1, \\ 0, & \theta \neq \theta_1, \end{cases}$$

où θ_1 est un point intérieur donné de Θ. Pour $w_2 = w_1(\theta) \equiv 1$ ceci correspond au problème de test de l'hypothèse simple $H_1 = \{\theta = \theta_1\}$ contre l'hypothèse complémentaire $H_2 = \{\theta \neq \theta_1\}$.

Pour trouver une décision bayésienne en appliquant le principe de Bayes, considérons un jeu ordinaire (D, Θ, w) et supposons que sur Θ est donnée une distribution $\mathbf{Q}$ telle que $q = \mathbf{Q}(\{\theta_1\}) > 0$ (nous nous plaçons dans l'approche totalement bayésienne). Posons $\mathbf{Q}_2 = \dfrac{\mathbf{Q} - q\mathbf{I}_{\theta_1}}{1 - q}$, où $\mathbf{I}_\theta$ est une distribution dégénérée concentrée au point θ. Alors

$$\tilde{w}(\delta_1, \mathbf{Q}) = (1 - q) \int w_1(t)\mathbf{Q}_2(dt), \quad \tilde{w}(\delta_2, \mathbf{Q}) = qw_2.$$

Ceci exprime que la stratégie bayésienne $\pi_{\mathbf{Q}}(\delta_2) = 1$ si

$$(1 - q) \int w_1(t)\mathbf{Q}_2(dt) > qw_2, \tag{3}$$

et $\pi_{\mathbf{Q}}(\delta_1) = 1$ si l'inégalité contraire a lieu. La relation (3) peut être mise sous la forme

$$\int w(t)\mathbf{Q}(dt) > 0,$$

où

$$w(t) = \begin{cases} w_1(t) & \text{si} \quad t \neq \theta_1, \\ -w_2 & \text{si} \quad t = \theta_1. \end{cases}$$

D'après le principe de Bayes, la décision bayésienne $\pi_Q(X)$ est de la forme $\pi_Q(X) = 1$ si

$$\int w(t) Q_X(dt) > 0,$$

où Q_X est une distribution *a posteriori*. Supposons que $\lambda(dt) = dt$ pour $t \neq \theta_1$, $\lambda(\{\theta_1\}) = 1$ et que la distribution Q_2 admet une densité $q_2(t)$ par rapport à la mesure de Lebesgue. Alors la distribution *a priori* Q admet une densité $q(t)$ par rapport à λ, égale à $(1 - q)q_2(t)$ pour $t \neq \theta_1$ et $q(t) = q$ pour $t = \theta_1$. Ceci exprime que la densité *a posteriori* par rapport à la mesure λ sera égale à

$$q(t|X) = \frac{f_t(X)q(t)}{f(x)},$$

$$f(X) = \int f_u(X)q(u)\lambda(du).$$

Donc, la décision bayésienne $\pi_Q(X)$ sera de la forme $\pi_Q(X) = 1$ si

$$(1 - q)\int w_1(t)f_t(X)q_2(t)dt > w_2 q f_{\theta_1}(X). \tag{4}$$

Le risque attaché à cette décision vaut

$$\tilde{W}(\pi_Q(\cdot), Q) = q w_2 \mathbf{P}_{\theta_1}(\pi_Q(X) = 1) + \\ + (1 - q)\int w_1(u)q_2(u)\mathbf{P}_u(\pi_Q(X) = 0)du.$$

En comparant ces relations avec le contenu du § 3.13, on constate que la région (4) d'acceptation de la décision δ_2 est ici de la même forme que la région $\Omega(c)$ dans (3.13.3) pour $c = w_2 q/(1 - q)$ et en remplaçant la fonction $q(t)$ par $w_1(t)q_2(t)$ dans (3.13.3). En d'autres termes

$$\pi_Q(X) = \begin{cases} 1 & \text{si} \quad r_{Q_2}(X) > c, \\ \gamma & \text{si} \quad r_{Q_2}(X) = c_1, \\ 0 & \text{si} \quad r_{Q_2}(X) < c, \end{cases} \tag{5}$$

où

$$r_{Q_2}(X) = \frac{\int w_1(t)q_2(t)f_t(X)dt}{f_{\theta_1}(X)}, \quad c = \frac{w_2 q}{1 - q}.$$

Suivant le § 3.13 on peut ensuite procéder comme suit. Dans l'ensemble des décisions bayésiennes (5) il faut choisir, en changeant le nombre q, la décision $\pi_Q(X)$ dont la valeur des « pertes de première espèce » soit fixe :

$$w_2\left[\mathbf{P}_{\theta_1}(\pi_Q(X) = 1) + \gamma \mathbf{P}_{\theta_1}(\pi_Q(X) = \gamma)\right] = \alpha.$$

De toutes les décisions $\pi(X)$ telles que

$$\alpha_1(\pi) = w_2 \mathbf{E}_{\theta_1}\pi(X) \leqslant \alpha, \tag{6}$$

la décision $\pi_Q(X)$ sera celle qui minimisera les « pertes de deuxième espèce »

$$\alpha_2(\pi) = \int w_1(u)q_2(u)\mathbf{E}_u(1 - \pi(X))du. \tag{7}$$

Ceci est la conséquence directe du fait que la décision π_Q est bayésienne. La comparaison des valeurs (6) et (7) aux probabilités d'erreur de première et de deuxième espèce (3.13.4) montre que nous avons de nouveau affaire à des distinctions insignifiantes dont la plus importante consiste à remplacer dans (3.13.4) la fonction $q(u)$ par la fonction $w_1(u)q_2(u)$. Les nombres c et γ de (5) sont définis à l'aide de α.

Ce qui vient d'être dit permet, en suivant exactement les raisonnements du § 3.13, d'énoncer les définitions et propositions suivantes.

DÉFINITION 1. On dit qu'une décision $\pi(X)$ appartient à la classe $\tilde{K}_\varepsilon$ (est de niveau asymptotique $1 - \epsilon$) si

$$\lim_{n\to\infty}\sup \mathbf{E}_{\theta_1}(x) \leqslant \epsilon .$$

Cette définition est pratiquement la même que la définition 3.13.1.

Montrons maintenant qu'en choisissant q convenablement, on peut faire en sorte que $\pi_Q \in \tilde{K}_\varepsilon$. Posons

$$r_{Q_2}(X) = \frac{\int w_1(t)q_2(t)f_t(X)dt}{f_{\theta_1}(X)} = \left(\frac{2\pi}{n}\right)^{k/2} \frac{w_1(\theta_1)q_2(\theta_1)}{\sqrt{|I|}} e^{T(X)},$$

où $I = I(\theta_1)$ est la matrice d'information de Fisher au point θ_1. Supposons par ailleurs que les conditions (RR) sont remplies, que θ_1 est un point intérieur de Θ et que la fonction $w_1(t)q_2(t)$ est continue au point θ_1 et strictement positive

$$c = \left(\frac{2\pi}{n}\right)^{k/2} \frac{w_1(\theta_1)q_2(\theta_1)}{\sqrt{|I|}} e^{z}. \tag{8}$$

Dans ces conditions, en vertu du lemme 3.13.1 on obtient pour la fonction $p_t(c) = \mathbf{P}_t(r_{Q_2}(X) > c)$

$$p_{\theta_1}(c) = \mathbf{P}_{\theta_1}(T(X) > z) \to \mathbf{H}_k(]2z, \infty[).$$

Donc, en posant $q = c/(c + w_2)$, où c est défini dans (8), $z = h_\varepsilon/2$, h_ε est le quantile d'ordre $1 - \varepsilon$ de la distribution du χ^2 à k degrés de liberté, on obtient

$$\lim_{n\to\infty} p_{\theta_1}\left(\frac{w_2 q}{1 - q}\right) = \epsilon ,$$

et par suite $\pi_Q(X) \in \tilde{K}_\varepsilon$.

DÉFINITION 2. On dit qu'une décision $\pi(X)$ est *asymptotiquement bayésienne dans* $\tilde{K}_\varepsilon$ pour une distribution *a priori* **Q** donnée si $\pi_Q \in \tilde{K}_\varepsilon$ et

$$\operatorname*{lim.sup}_{n\to\infty} \frac{\alpha_2(\pi)}{\alpha_2(\pi_Q)} = 1.$$

THÉORÈME 1. *Si les conditions (RR) sont satisfaites et* θ_1 *est un point intérieur de* Θ, *il existe alors dans* $\tilde{K}_\varepsilon$ *une décision asymptotiquement bayésienne* $\hat{\pi}(X)$, *la même pour toutes les distributions* $\mathbf{Q}_2$ *et toutes les fonctions* $w_1(t)$ *telles que la fonction* $w_1(t)q_2(t)$ *est continue, strictement positive en* θ_1 *et bornée sur* Θ. *Le test* $\hat{\pi}$ *est défini par la relation*

$$\hat{\pi}(X) = 1 \quad si \quad \frac{f_{\theta^*}(X)}{f_{\theta_1}(X)} > e^{k_\varepsilon\sqrt{2}}. \tag{9}$$

Ce théorème se prouve exactement comme le théorème 3.13.1 au changement près de la fonction $q(t)$ en la fonction $w_1(t)q_2(t)$. Le théorème 3.13.1 permet de déterminer aussi la valeur des « pertes de deuxième espèce » (cf. (7)) du test $\hat{\pi}$.

Le test (9) n'est autre qu'un test du rapport de vraisemblance.

§ 8. Décisions asymptotiquement optimales avec une fonction de perte arbitraire dans le cas d'hypothèses proches

Dans ce paragraphe on généralise les résultats du § 3.14 au cas d'une fonction de perte arbitraire. Cette généralisation sera plus consistante que dans le paragraphe précédent, puisque la fonction de perte dépendra de n (comparer avec le § 6).

Soit $(\mathscr{D}, \Theta, W)$ un jeu statistique dans lequel $\Theta \subset R^k$, $D = \{\delta_1, \delta_2\}$, $w(\delta_i, \theta) = w_i(\theta)$, où $w_i(\theta) = 0$ pour $\theta \in \Theta_i$, $i = 1, 2$, $\Theta_1 \cap \Theta_2 = \varnothing$.

Si $w_i(\theta) = 1$ pour $\theta \notin \Theta_i$, on obtient un problème de test des hypothèses $H_i = \{\theta \in \Theta_i\}$, $i = 1, 2$.

Trouvons une stratégie bayésienne pour le jeu (D, Θ, w). Soient $\mathbf{Q}_i$ des distributions sur Θ_i

$$\mathbf{Q} = q_1\mathbf{Q}_1 + q_2\mathbf{Q}_2, \quad q_1 + q_2 = 1.$$

Il est alors évident que

$$\tilde{w}(\delta_i, \mathbf{Q}) = \int w_i(t)\mathbf{Q}(dt) \quad \text{et} \quad \pi_Q(\delta_2) = 1,$$

si

$$\int w_2(t)\mathbf{Q}(dt) < \int w_1(t)\mathbf{Q}(dt),$$

ou si

$$q_1\int w_2(t)\mathbf{Q}_1(dt) < q_2\int w_1(t)\mathbf{Q}_2(dt).$$

En vertu donc du principe de Bayes, la décision bayésienne $\pi_Q(X)$ sera de la forme $\pi_Q(X) = 1$ si

$$\int w_2(t)Q_X(dt) < \int w_1(t)Q_X(dt). \tag{1}$$

Supposons que les distributions Q_i àdmettent les densités $q_i(t)$, $i = 1, 2$, par rapport à une mesure λ. Alors Q et la distribution *a posteriori* Q_X auront des densités égales respectivement à $q(t) = q_1q_1(t) + q_2q_2(t)$ et

$$q(t|X) = \frac{q(t)f_t(X)}{f(X)}, \quad f(X) = \int q(u)f_u(X)\lambda(du).$$

Ceci exprime que la relation (1) peut être mise sous la forme

$$q_1 \int_{\Theta_1} w_2(t)q_1(t)f_t(X)\lambda(dt) < q_2 \int_{\Theta_2} w_1(t)q_2(t)f_t(X)\lambda(dt). \tag{2}$$

Le risque attaché à la décision bayésienne $\pi_Q(X)$ vaut

$$W(\pi_Q(\cdot), \theta) = w_1(\theta)\mathbf{E}_\theta\pi_Q(X) + w_2(\theta)(1 - \mathbf{E}_\theta\pi_Q(X)),$$
$$\bar{W}(\pi_Q(\cdot), Q) = \int W(\pi_Q(\cdot), t)q(t)\lambda(dt).$$

Passons maintenant à l'examen *d'alternatives proches*. Soit θ_1 une valeur quelconque fixe du paramètre θ. Comme dans le § 3.14, on admettra que les ensembles Θ_i sont de la forme

$$\Theta_i = \theta_1 + \Gamma_i/\sqrt{n}, \tag{3}$$

où Γ_i sont indépendants de n. A propos de Q_i on admettra qu'elles sont induites par des distributions Π_i concentrées sur Γ_i et indépendantes de n. Si les ensembles Γ_i sont bornés, les stratégies θ sont situées dans un $1/\sqrt{n}$-voisinage du point θ_1. Si donc $w_1(t)$ et $w_2(t)$ sont continues, $w_i(t) > c > > 0$, $i = 1, 2$, respectivement sur les ensembles Θ_2 et Θ_1, le jeu statistique $(\mathscr{D}, \Theta, W)$ caractérisé par une telle fonction de perte possédera pratiquement les mêmes propriétés que le jeu de fonction de perte $w_i(t) = 1$, $t \notin \Theta_i$, étudié dans les §§ 3.14 et 3.15.

Nous envisageons ici une généralisation plus consistante identique à celle qui a été conduite dans le § 6. On admettra que la fonction de perte $w(\delta_i, \theta) = w_i(\theta)$ dépend de n, si bien que

$$w_i(\theta) = w_{i,n}(\theta) = v_i(\sqrt{n}(\theta - \theta_1)), \tag{4}$$

où $v_i(t)$ sont des fonctions mesurables bornées indépendantes de n.

Suivant le § 3.14, on appellera *problème* A le problème qui consiste à trouver à partir d'un échantillon $X \in \mathbf{P}_\theta$ une décision du jeu $(\mathscr{D}, \Theta, W)$ décrit plus haut. Si les relations (3) et (4) ont lieu, on dira que le problème A est un problème de test d'hypothèses proches avec des fonctions de perte $v_i(t)$.

Considérons maintenant un autre jeu statistique $(\mathscr{D}_B, \Gamma, V)$ relatif à un échantillon $Y \in \Phi_{\gamma, I^{-1}}$ de taille un, où $I = I(\theta_1)$ est la matrice d'information de Fisher pour la famille $\{\mathbf{P}_\theta\}$ au point θ_1. Les éléments de ce jeu sont : l'ensemble des décisions $D_B = \{d_1, d_2\}$ et l'ensemble des stratégies de la nature $\Gamma = \Gamma_1 \cup \Gamma_2$. La fonction de perte $v(d, \gamma) : D_B \times \Gamma \to R$ est définie par les relations

$$v(d_i, \gamma) = v_i(\gamma). \ v_i(\gamma) = 0 \text{ si } \gamma \in \Gamma_i.$$

Donc, dans ce jeu, $\mathscr{D}_B$ est la classe de toutes les décisions $d(Y) : \mathscr{Y} = R^k \to D_B$,

$$V(d(\cdot), \gamma) = v_1(\gamma)\Phi_{\gamma, I^{-1}}(d(Y) = d_1) + v_2(\gamma)\Phi_{\gamma, I^{-1}}(d(Y) = d_2)$$

(l'un des termes du second membre est nul). On note de façon analogue les pertes attachées aux stratégies randomisées $\pi(Y)$ en termes de $\mathbf{E}\pi(Y)$, $Y \in \Phi_{\gamma, I^{-1}}$. Ce problème sera appelé *problème* B.

Les problèmes A et B sont reliés par la même relation que les problèmes homologues du § 3.14. Soit $\pi(Y)$ une décision du problème B optimale dans un sens ou dans l'autre (bayésienne, minimax), et soit $\hat{\theta}^*$ l'estimateur du maximum de vraisemblance dans le problème A, $\gamma^* = (\hat{\theta}^* - \theta_1)\sqrt{n}$. La décision $\pi(\gamma^*)$ sera alors une décision asymptotiquement optimale (dans le même sens) du problème A.

Le « critère limite d'optimalité » formulé permet de ramener le problème A à un problème B plus simple.

Pour donner un sens plus précis à ce qui vient d' être dit, considérons les définitions suivantes. Soient données des distributions Π_i sur Γ_i. Posons $\Pi = q_1\Pi_1 + q_2\Pi_2$, $q_1 + q_2 = 1$, et désignons par $\mathbf{Q}$ la distribution induite sur Θ par Π et la transformation $\theta = \theta_1 + \gamma/\sqrt{n}$.

DÉFINITION 1. On dit qu'une décision $\pi_1(X)$ est *asymptotiquement bayésienne* si

$$\limsup_{n \to \infty}[\tilde{W}(\pi_1(\cdot), \mathbf{Q}) - \tilde{W}(\pi_\mathbf{Q}(\cdot), \mathbf{Q})] \leqslant 0.$$

Comme précédemment

$$\tilde{W}(\pi(\cdot), \theta) = w_1(\theta)\mathbf{E}_\theta\pi(X) + w_2(\theta)(1 - \mathbf{E}_\theta\pi(X)),$$
$$\tilde{W}(\pi(\cdot), \theta) = \int(\tilde{W}/\pi(\cdot), t)\mathbf{Q}(dt),$$

où $\pi_\mathbf{Q}$ est une décision bayésienne.

DÉFINITION 2. On dit qu'une décision $\pi_1(X)$ est *asymptotiquement minimax* si pour toute autre décision $\pi(X)$ on a

$$\limsup_{n \to \infty}\left[\sup_{\theta \in \Theta}\tilde{W}(\pi_1(\cdot), \theta) - \sup_{\theta \in \Theta}\tilde{W}(\pi(\cdot), \theta)\right] \leqslant 0.$$

On aurait pu comparer ici π_1 uniquement à une stratégie minimax $\bar{\pi}$ (comparer avec la définition 1).

Nous aurions pu comme dans le § 3.14 envisager aussi des décisions asymptotiquement bayésiennes et asymptotiquement minimax dans la classe $\tilde{K}_\epsilon$ des décisions à « pertes de première espèce » asymptotiques fixes

$$\epsilon = \limsup_{n\to\infty} \sup_{\theta\in\Theta_1} w_1(\theta)\mathbf{E}_\theta\pi(X).$$

Pour obtenir les résultats respectifs, il suffira de comparer le contenu de ce paragraphe à celui du § 3.14.

Désignons par $\pi_\Pi(Y)$ la décision bayésienne du jeu $(\mathscr{D}_B, \Gamma, V)$ (c'est-à-dire du problème B) associée à une distribution *a priori* Π et supposons pour simplifier que les ensembles Γ_i sont bornés.

THÉORÈME 1. *Supposons qu'au voisinage d'un point* θ_1 *les conditions* (*RR*) *sont satisfaites et que les fonctions* v_i *et la distribution* Π_i *sont telles que* $0 < \int v_1(u)\Pi_2(du) < \infty$, $0 < \int v_2(u)\Pi_1(du) < \infty$. *Alors dans les notations introduites, le test*

$$\pi_1(X) = \pi_\Pi(\gamma^*), \ \gamma^* = (\hat{\theta}^* - \theta_1)\sqrt{n},$$

sera la décision asymptotiquement bayésienne du jeu $(\mathscr{D}, \Theta, W)$ (*c'est-à-dire du problème* A) *associée à la distribution a priori* $\mathbf{Q}$.

THÉORÈME 2. *Supposons qu'au voisinage de* θ_1 *sont satisfaites les conditions* (*RR*) *et que dans le problème* B *existent la décision minimax* $\bar{\pi}(Y)$ *et la distribution* $\bar{\Pi}$ *la plus défavorable correspondante. Le test* $\pi_1(X) = \bar{\pi}(\gamma^*)$ *sera alors une décision asymptotiquement minimax du problème* A.

REMARQUE 1. En vertu des théorèmes du § 3, les conditions d'existence de $\bar{\pi}$ et $\bar{\Pi}$ seront réunies si v_i sont des fonctions continues.

DÉMONSTRATION du théorème 1. Elle est calquée sur celle du théorème 3.14.1. De (2) il s'ensuit que la décision bayésienne π_Q est de la forme $\pi_Q(X) = 1$ si

$$\frac{\int w_1(t)q_2(t)f_t(X)\lambda(dt)}{\int w_2(t)q_1(t)f_t(X)\lambda(dt)} > \frac{q_1}{q_2}. \tag{5}$$

En posant $Z_1(t) = \dfrac{f_{\theta_1+t}(X)}{f_{\theta_1}(X)}$ et puisque

$$q_i(t)\lambda(dt) = \mathbf{Q}_i(dt), \ \mathbf{Q}_i(\theta_1 + du/\sqrt{n}) = \Pi_i(du),$$
$$w_i(\theta_1 + u/\sqrt{n}) = v_i(u),$$

on peut par le changement $t = \theta_1 + u/\sqrt{n}$ ramener l'inégalité (5) à la forme

$$\frac{\int v_1(u)Z_1(u/\sqrt{n})\Pi_2(du)}{\int v_2(u)Z_1(u/\sqrt{n})\Pi_1(du)} = \frac{\int Z_1(u/\sqrt{n})\Pi_2'(du)}{\int Z_1(u/\sqrt{n})\Pi_1'(du)} > c, \quad c = \frac{q_1}{q_2}, \quad (6)$$

où les distributions généralisées $\Pi_i'(A) = \int_A v_{i+1}(u)\Pi_i(du)$ ($v_3(u) \equiv v_1(u)$, $i = 1, 2$) peuvent être transformées en mesures de probabilité par une renormalisation en introduisant les distributions $\Pi_i''(A) = \Pi_i'(A)/\Pi_i'(\Gamma_i)$ (par hypothèse $0 < \Pi_i'(\Gamma_i) < \infty$). Nous obtenons alors en qualité de (5) une inégalité exactement de la même forme que dans le § 3.14.

La suite de la démonstration est la même que dans le § 3.14 à quelques simplifications près. Nous la laissons au soin du lecteur. Signalons qu'elle s'appuie sur la convergence uniforme en γ

$$\tilde{W}(\pi_Q(\cdot), \theta) \to \tilde{V}(\pi_\Pi(\cdot), \gamma), \; \tilde{W}(\pi_1(\cdot), \theta) \to \tilde{V}(\pi_\Pi(\cdot), \gamma), \qquad (7)$$

où $\pi_1(X) = \pi_\Pi(\gamma^*)$ et $\theta = \theta_1 + \gamma/\sqrt{n}$. ◄

Pour prouver le théorème 2 nous aurons besoin du

LEMME 1. *Soient* Q *une distribution a priori et* π_1 *la décision asymptotiquement bayésienne telle que*

$$\limsup_{n\to\infty} \tilde{W}(\pi_1(\cdot), Q) = c, \; \limsup_{n\to\infty} \sup_{\theta\in\Theta} \tilde{W}(\pi(\cdot), \theta) \leqslant c. \qquad (8)$$

Alors π_1 *est une décision asymptotiquement minimax.*

DÉMONSTRATION. Désignons comme précédemment par π_Q la décision bayésienne. Pour toute décision π on a alors

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \tilde{W}(\pi, \theta) \geqslant \limsup_{n\to\infty} \tilde{W}(\pi, Q) \geqslant$$
$$\geqslant \limsup_{n\to\infty} \tilde{W}(\pi_Q, Q) \geqslant \limsup_{n\to\infty} \tilde{W}(\pi_1, Q) =$$
$$= c \geqslant \limsup_{n\to\infty} \sup_{\theta\in\Theta} \tilde{W}(\pi_1, \theta). \; ◄$$

DÉMONSTRATION du théorème 2. Soit $\overline{\Pi}$ une distribution la plus défavorable sur Γ, de sorte que $\overline{\pi}(Y) = \pi_{\overline{\Pi}}(Y)$ est une décision minimax du jeu $(\mathcal{Y}_B, \Gamma, V)$. Le théorème 1 nous dit alors que $\pi_1(X) = \pi_{\overline{\Pi}}(\gamma^*)$ sera une décision asymptotiquement bayésienne pour la distribution $\overline{Q}$ associée à $\overline{\Pi}$ et

pour prouver le théorème il suffit de s'assurer que $\overline{Q}$ et π_1 satisfont les conditions du lemme 1.

Désignons par $N_{\overline{\Pi}}$ le support de la distribution $\overline{\Pi}$. Les théorèmes du § 3 nous donnent alors

$$\begin{gathered} \tilde{V}(\pi_{\overline{\Pi}}(\cdot), \gamma) = c, \ \gamma \in N_{\overline{\Pi}}, \\ \sup_{\gamma \in \Gamma} \tilde{V}(\pi_{\overline{\Pi}}(\cdot) \leqslant c. \end{gathered} \tag{9}$$

Mais si $\theta = \theta_1 + \gamma/\sqrt{n}$, on a la convergence $\tilde{W}(\pi_1(\cdot), \theta) \to \tilde{V}(\pi_{\overline{\Pi}}(\cdot), \gamma)$ uniformément en γ (cf. (7)). Ceci et (9) entraînent (8). ◄

ANNEXE I

THÉORÈMES DE TYPE GLIVENKO-CANTELLI

Dans cette Annexe on prouvera des propositions qui entraîneront les théorèmes 1.4.1 et 1.4.2. On se servira sans explications des notations du paragraphe 1.4 dans lesquelles ces théorèmes sont formulés. Démontrons tout d'abord une version générale auxiliaire du théorème de Glivenko-Cantelli.

DÉFINITION 1. On dira qu'une classe $\mathfrak{K}$ d'ensembles de $\mathfrak{B}_{\mathcal{X}} = \mathfrak{B}^m$ est *finiment-approximable* (par rapport à une distribution $\mathbf{P}$) si pour tout $\epsilon > 0$ il existe une autre classe d'ensembles $\mathfrak{S}(\epsilon)$, composé d'un nombre fini $N = N(\epsilon)$ d'éléments $S_1, \ldots, S_N$, $S_i \in \mathfrak{B}^m$, telle que pour tout $B \in \mathfrak{K}$ on peut exhiber des ensembles A_1 et A_2 de $\mathfrak{S}(\epsilon)$ jouissant des propriétés suivantes

$$A_1 \subset B \subset A_2, \qquad \mathbf{P}(A_2 - A_1) < \epsilon. \tag{1}$$

Définissons *l'addition, la multiplication* et la *complémentation* sur les classes $\mathfrak{K}$. On appellera classes $\mathfrak{K}_1 + \mathfrak{K}_2$ et $\mathfrak{K}_1\mathfrak{K}_2$ respectivement les classes d'ensembles de la forme $A \cup B$ et $A \cap B$, où $A \in \mathfrak{K}_1$ et $B \in \mathfrak{K}_2$. On appellera complémentaire $\overline{\mathfrak{K}}$ la classe des complémentaires $\overline{A}$, $A \in \mathfrak{K}$.

THÉORÈME 1. 1) *Supposons que* $X_n = [X_\infty]_n$, $X_\infty \in \mathbf{P}$ *et que* $\mathfrak{K}$ *est une classe finiment-approximable. Alors*

$$\sup_{B \in \mathfrak{K}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{p.s.}}{\to} 0. \tag{2}$$

2) *L'ensemble des classes finiment-approximables est stable pour les opérations définies.*

DÉMONSTRATION. La première proposition s'établit à l'aide des raisonnements utilisés pour le cas scalaire dans le théorème 1.2.2. Pour $B \in \mathfrak{K}$ et $\varepsilon > 0$ donnés, on peut exhiber un $N = N(\varepsilon)$ et des ensembles A_1 et A_2 doués de la propriété (1). On a pour ces ensembles

$$\mathbf{P}_n^*(B) - \mathbf{P}(B) \leqslant \mathbf{P}_n^*(A_2) - \mathbf{P}(A_1) < \mathbf{P}_n^*(A_2) - \mathbf{P}(A_2) + \epsilon,$$

$$\mathbf{P}_n^*(B) - \mathbf{P}(B) \geqslant \mathbf{P}_n^*(A_1) - \mathbf{P}(A_2) > \mathbf{P}_n^*(A_1) - \mathbf{P}(A_1) - \epsilon.$$

Donc

$$\bigcap_{k=1}^{N} \{|\mathbf{P}_n^*(S_k) - \mathbf{P}(S_k)| < \epsilon\} \subset \left\{\sup_{B \in \mathfrak{K}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| < 2\epsilon\right\},$$

où $S_1, \ldots, S_N$ sont des éléments de $\mathfrak{S}(\epsilon)$. Comme $\mathbf{P}_n^*(S_k) \underset{\text{p.s.}}{\to} \mathbf{P}(S_k)$, on en déduit sans peine (2) (comparer avec la démonstration du théorème 1.2.2A).

La deuxième assertion du théorème 1 est presque évidente. Soit donné $\epsilon > 0$ et supposons que $\mathfrak{S}_1(\epsilon_1)$ et $\mathfrak{S}_2(\epsilon_2)$ sont des classes approximant $\mathfrak{K}_1$ et $\mathfrak{K}_2$ respectivement. Supposons par ail-

leurs que A et B sont des ensembles quelconques de $\mathfrak{K}_1$ et $\mathfrak{K}_2$. Les relations $\epsilon_1 + \epsilon_2 = \epsilon$

$$A_1 \subset A \subset A_2, \mathbf{P}(A_2 - A_1) < \epsilon_1 \quad (A_i \in \mathfrak{S}_1(\epsilon_1)),$$
$$B_1 \subset B \subset B_2, \mathbf{P}(B_2 - B_1) < \epsilon_2 \quad (B_i \in \mathfrak{S}_2(\epsilon_2)),$$

entraînent

$$A_1B_1 \subset AB \subset A_2B_2,$$
$$A_2B_2 - A_1B_1 \subset (A_2 - A_1) \cup (B_2 - B_1),$$
$$\mathbf{P}(A_2B_2 - A_1B_1) \leqslant \epsilon.$$

Donc, la classe $\mathfrak{K}_1\mathfrak{K}_2$ est finiment-approximable. La somme $\mathfrak{K}_1 + \mathfrak{K}_2$ et le complémentaire $\overline{\mathfrak{K}}$ se traitent de façon analogue. ◄

Corollaire 1. *Si $\mathscr{X} = R^m$, $X_n = [X_\infty]_n \in F$, alors*

$$\sup_t |F_n^*(t) - F(t)| \underset{p.s.}{\to} 0$$

lorsque $n \to \infty$, où $F_n^*(t)$ est une fonction de répartition empirique.

Démonstration. On voit sur la démonstration du théorème 1.2.2A que les classes $\mathfrak{K}_j = \{y \in R^m : y_j < t_j\}$, $-\infty < t_j < \infty$, sont finiment-approximables pour tout $j = 1, \ldots, m$. Pour système $\mathfrak{S}(\epsilon)$ il suffit de prendre des semi-espaces $\{y_j < z_k\}$ et $\{y_j \leqslant z_k\}$, $k = 1, \ldots, N$, où z_k sont définis dans (1.2.6).

La classe des angles $\mathfrak{K} = \mathfrak{K}_1\mathfrak{K}_2 \ldots \mathfrak{K}_m$ sera aussi finiment-approximable en vertu de la deuxième proposition du théorème 1. Reste maintenant à se servir de la première proposition du théorème 1. ◄

Le corollaire 1 n'est autre que le théorème 1.4.1.

Considérons maintenant les classes $\mathfrak{K}$ satisfaisant la condition (Γ) suivante. Soit K_M le cube

$$K_M = \left\{ y = (y_1, \ldots, y_m) : \max_{1 \leqslant k \leqslant m} |y_k| \leqslant M \right\}.$$

(Γ) *Les ensembles $B \in \mathfrak{K}$ jouissent de la propriété suivante : tout ϵ-voisinage Γ_B^ϵ de la frontière $\Gamma_B = \partial(B \cap K_M)$ possède une mesure de Lebesgue (un volume) $\mu(\Gamma_B^\epsilon) \leqslant \varphi(\epsilon, M)$, où φ ne dépend que de ses arguments et $\varphi(\epsilon, M) \to 0$ lorsque $\epsilon \to 0$ pour tout M.*

Théorème 2. *Si $\mathscr{X} = R^m$, $X \in \mathbf{P}$, où $\mathbf{P}$ est une distribution absolument continue par rapport à la mesure de Lebesgue, alors toute classe $\mathfrak{K}$ satisfaisant la condition (Γ) est finiment-approximable et, par suite, est justiciable de (2).*

Démonstration. Remarquons tout d'abord que le problème sur l'espace R^m peut être ramené à un problème sur le cube K_M au sens suivant. Supposons que pour chaque M fixe, il existe une classe $\mathfrak{S}$ de sous-ensembles de K_M telle que pour tout $B' \in \mathfrak{K}$ et $B = B' \cap K_M$ soit réalisée (1). Dans ces conditions $\mathfrak{K}$ sera finiment-approximable. En effet, pour $\varepsilon > 0$ choisi dans (1), trouvons un $M = M(\epsilon)$, tel que $\mathbf{P}(K_M) \geqslant 1 - \epsilon$ et posons $A_1' = A_1$, $A_2' = A_2 \cup \overline{K}_M$, où A_i sont les ensembles de (1), $\overline{K}_M$ le complémentaire de K_M. Il est alors évident que

$$A_1' \subset B' \subset A_2', \quad \mathbf{P}(A_2' - A_1') \leqslant 2\epsilon.$$

Nous pouvons donc considérer que $\mathbf{P}(K_M) = 1$ et que $\mathfrak{K}$ est composé de sous-ensembles de K_M.

Prenons pour $\mathfrak{S}$ les figures A_j constituées des diverses réunions des cubes fermés d'arêtes δ et de sommets

$$(j_1\delta, \ldots, j_m\delta), \quad -M/\delta < j_k < M/\delta, \quad k = 1, \ldots, m,$$

(pour simplifier on peut admettre que δ est aliquote de M). Définissons les ensembles A_1 et A_2 respectivement comme les réunions de tous les cubes appartenant à B et l'intersectant. Il est évident que

$$A_1 \subset B \subset A_2,$$

$$\mu(A_2 - A_1) \leqslant \varphi(\Gamma_B^{2\delta\sqrt{m}}) \leqslant \sigma(2\delta\sqrt{m}, M).$$

Le membre de droite de cette inégalité peut être rendu aussi petit que l'on veut par un choix convenable de δ.

Par ailleurs $\mathbf{P}$ est absolument continue par rapport à μ. Donc, pour ε donné, on peut trouver un $\gamma = \gamma(\epsilon)$ tel que $\sup\limits_{\mu(A)<\gamma} \mathbf{P}(A) < \epsilon$. Si maintenant l'on choisit δ de telle sorte que $\varphi(2\delta\sqrt{m}, M) < \gamma$, on obtient

$$\mathbf{P}(A_2 - A_1) < \epsilon. \quad ◄$$

COROLLAIRE 2. *La classe $\mathfrak{C}$ de tous les ensembles convexes est finiment-approximable et, par suite, pour les distributions* $\mathbf{P}$ *absolument continues on a*

$$\sup_{B\in\mathfrak{C}} |\mathbf{P}_n^*(B) - \mathbf{P}(B)| \underset{\text{p.s.}}{\rightarrow} 0.$$

En effet la plus grande « aire » d'un ensemble convexe dans K_M est égale à $2m(2M)^{m-1}$ (qui est l« aire » de K_M), et le volume maximal $\mu((\partial K_M)^\epsilon)$ d'un ϵ-voisinage de ∂K_M est au plus égal à $2\epsilon\, 2m(2M)^{m-1}$. Ceci exprime que la condition (Γ) est satisfaite. ◄

Le corollaire 2 coïncide avec le théorème 1.4.2. Dans le § 1.4 on trouvera une remarque relative à l'importance de la condition de continuité absolue de $\mathbf{P}$.

Il est immédiat de voir que la condition (Γ) sera également remplie pour les classes d'ensembles non convexes à frontières suffisamment différentiables.

ANNEXE II

THÉORÈME LIMITE FONCTIONNEL POUR PROCESSUS EMPIRIQUES

On se propose de prouver l'assertion suivante (théorème 1.6.3). Supposons que

$$w^n(t) = \sqrt{n}(F_n^*(t) - t)$$

est le processus empirique défini dans le § 1.6 et que $w^\circ(t)$ est un pont brownien.

THÉORÈME 1. *Si f est une fonctionnelle mesurable de $D(0, 1)$ dans R, continue sur l'espace $C(0, 1)$ pour une métrique uniforme, alors pour $n \to \infty$*

$$f(w^n) \Rightarrow f(w^\circ).$$

La démonstration de ce théorème passe par celle des deux lemmes suivants.

LEMME 1. *Les distributions finidimensionnelles des processus w^n convergent faiblement pour $n \to \infty$ vers les distributions correspondantes du processus w°.*

DÉMONSTRATION. Considérons les vecteurs aléatoires $(m + 1)$-dimensionnels

$$\mathbf{w}^n = (\Delta_0 w^n, \ldots, \Delta_m w^n),$$

où $\Delta_j g$ désigne, comme dans le § 1.6, les écarts

$$\Delta_j g = g(t_{j+1}) - g(t_j),$$

$$t_{j+1} > t_j, \; j = 0, \ldots, m, \; t_0 = 0, \; t_{m+1} = 1.$$

Désignons par $\mathbf{w}^\circ$ le vecteur analogue pour le processus $w^\circ(t)$. Pour prouver ce lemme, il suffit en vertu du deuxième théorème de continuité de montrer que $\mathbf{w}^n \Rightarrow \mathbf{w}^\circ$.

Trouvons les fonctions caractéristiques de $\mathbf{w}^n$ et $\mathbf{w}^\circ$. Pour le vecteur $u = (u_0, \ldots u_m)$ on a

$$\mathbf{E}e^{i\mathbf{w}^\circ u^T} = \mathbf{E}\exp\left\{i\sum_{j=0}^{m} u_j\Delta_j w^\circ\right\} = \mathbf{E}\exp\left\{i\sum_{j=0}^{m} u_j(\Delta_j w - w(1)\Delta_j)\right\},$$

où $\Delta_j = t_{j+1} - t_j$, $j = 0, \ldots, m$, et $w(t)$ est un processus wienérien standard.

Représentons l'exposant de l'exponentielle par une somme de variables indépendantes. En posant pour simplifier $\sum_{j=0}^{m} u_j\Delta_j = U$, on obtient

$$\sum_{j=0}^{m} u_j(\Delta_j w - w(1)\,\Delta_j) = \sum_{j=0}^{m} (u_j - U)\Delta_j w.$$

Comme $Ee^{iuw(t)} = e^{-u^2t/2}$, il vient

$$Ee^{iw^*u^T} = \exp\left\{-\frac{1}{2}\sum_{j=0}^{m}(u_j - U)^2\Delta_j\right\} = \exp\left\{-\frac{1}{2}\left(\sum_{j=0}^{m}u_j^2\Delta_j - U^2\right)\right\}. \qquad (1)$$

Considérons maintenant la quantité $Ee^{iw^nu^T}$. Supposons comme précédemment (cf. § 1.6) que

$$\pi_n(t) = nF_n^*(t).$$

On sait alors que (cf. (1.6.1))

$$\mathrm{P}(\Delta_0\pi_n = k_0, \ldots, \Delta_m\pi_n = k_m) = \frac{n!}{k_0! \ldots k_m!}\Delta_0^{k_0} \ldots \Delta_m^{k_m}.$$

Le second membre est composé des termes du développement du polynôme $(\Delta_0 + \ldots + \Delta_m)^n$. En se servant de ce fait, on obtient

$$Ee^{i\sum_{j=0}^{m}u_j\Delta_j\pi_n} = \sum\frac{n!}{k_0! \ldots k_m!}(e^{iu_0}\Delta_0)^{k_0} \ldots (e^{iu_m}\Delta_m) =$$

$$= (e^{iu_0}\Delta_0 + \ldots + e^{iu_m}\Delta_m)^n.$$

Puisque $\Delta_jw^n = \sqrt{n}(F_n^*(t_{j+1}) - F_n^*(t_j) - \Delta_j) = (\Delta_j\pi_n - n\Delta_j)/\sqrt{n}$, il vient

$$Ee^{iw^nu^T} = \exp\left\{-i\sum_{j=0}^{m}u_j\sqrt{n}\Delta_j\right\}E\exp\left\{\frac{i}{\sqrt{n}}\sum_{j=0}^{m}u_j\Delta_j\pi_n\right\} = e^{i\sqrt{n}U}\left(\sum_{j=0}^{m}e^{iu_j/\sqrt{n}}\Delta_j\right)^n.$$

D'où l'on déduit, grâce aux égalités

$$e^\alpha = 1 + \alpha + \alpha^2/2 + O(\alpha^3), \quad \ln(1 + \alpha) = \alpha - \alpha^2/2 + O(\alpha^3)$$

que

$$\ln Ee^{iw^nu^T} = -iU\sqrt{n} + n\ln\left[1 - \sum_{j=0}^{m}(1 - e^{iu_j/\sqrt{n}})\Delta_j\right] =$$

$$= -iU\sqrt{n} + n\ln\left[1 + \sum_{j=0}^{m}\left(i\frac{u_j}{\sqrt{n}} - \frac{u_j^2}{2n} + O(n^{-3/2})\right)\Delta_j\right] =$$

$$= -iU\sqrt{n} + n\left[\frac{iU}{\sqrt{n}} - \frac{1}{2n}\sum_{j=0}^{m}u_j^2\Delta_j + \frac{U^2}{2n} + O(n^{-3/2})\right] =$$

$$= \frac{1}{2}\left[-\sum_{j=0}^{m}u_j^2\Delta_j + U^2\right] + O(n^{-1/2})$$

pour u fixe et $\alpha = o(1)$. En comparant avec (1) on voit que pour $n \to \infty$

$$Ee^{iw^nu^T} \to Ee^{iw^*u^T}. \qquad (2)$$

Reste à appliquer le théorème de continuité pour les fonctions caractéristiques des distributions multidimensionnelles (cf. [11]). ◄

LEMME 2. *Pour tout* $\epsilon > 0$

$$\limsup_{n\to\infty} \mathbf{P}(\omega_\Delta(\omega^n) > \epsilon) \to 0 \quad . \tag{3}$$

lorsque $\Delta \to 0$, *où* $\omega_\Delta(y)$ *est le module de continuité de la fonction* $y \in D(0, 1)$: $\omega_\Delta(y) = \sup_{\substack{0 \leqslant t_1 < t_2 \leqslant 1 \\ |t_1 - t_2| \leqslant \Delta}} |y(t_1) - y(t_2)|$.

DÉMONSTRATION. Sans perdre en généralité, on peut se borner aux seuls nombres dyadiques $\Delta = 2^{-l}$. Pour $m > l$ on a

$$\omega_\Delta(w^n) \leqslant \omega_\Delta^{[m]} + 2 \max_{k \leqslant 2^m} \omega\left(\frac{k-1}{2^m}, \frac{k}{2^m}\right),$$

où

$$\omega_\Delta^{[m]} = \max_{\left|\frac{k-j}{2^m}\right| \leqslant \Delta} \left| w^n\left(\frac{k}{2^m}\right) - w^n\left(\frac{j}{2^m}\right)\right|,$$

$$\omega\left(\frac{k-1}{2^m}, \frac{k}{2^m}\right) = \sup_{\frac{k-1}{2^m} \leqslant u < \frac{k}{2^m}} \left| w^n\left(\frac{k}{2^m}\right) - w^n(u)\right|.$$

Pour prouver (3), considérons

$$\mathbf{P}(\omega_\Delta(w^n) \geqslant 3\epsilon) \leqslant \mathbf{P}(\omega_\Delta^{[m]} \geqslant \epsilon) + \mathbf{P}\left(\bigcup_{k=1}^{2^m}\left\{\omega\left(\frac{k-1}{2^m}, \frac{k}{2^m}\right) \geqslant \epsilon\right\}\right). \tag{4}$$

Voyons le premier terme. Il est immédiat de voir que pour $l > 3$ l'événement

$$\bigcap_{r=l}^{m}\bigcap_{k=1}^{2^r}\left\{\left|w^n\left(\frac{k}{2^r}\right) - w^n\left(\frac{k-1}{2^r}\right)\right| < \frac{\epsilon}{r^2}\right\}$$

entraîne $\{\omega_\Delta^{[m]} < \epsilon\}$. Vu que l'inégalité inverse est réalisée pour les événements complémentaires, il vient

$$\mathbf{P}(\omega_\Delta^{[m]} \geqslant \epsilon) \leqslant \mathbf{P}\left(\bigcup_{r=l}^{m}\bigcup_{k=1}^{2^r}\left\{\left|w^n\left(\frac{k}{2^r}\right) - w^n\left(\frac{k-1}{2^r}\right)\right| \geqslant \frac{\epsilon}{r^2}\right\}\right). \tag{5}$$

Mais $w^n\left(\frac{k}{2^r}\right) - w^n\left(\frac{k-1}{2^r}\right) = \sqrt{n}\left(F_n^*\left(\frac{k}{2^r}\right) - F_n^*\left(\frac{k-1}{2^r}\right) - \frac{1}{2^r}\right)$, où $n\left(F_n^*\left(\frac{k}{2^r}\right) - F_n^*\left(\frac{k-1}{2^r}\right)\right)$ est la fréquence d'accès des éléments de l'échantillon dans

un intervalle de longueur 2^{-r}. En d'autres termes, ceci est la somme S_n des variables aléatoires dans une série de Bernoulli de n épreuves dont la probabilité de l'issue 1 est égale à $p = 2^{-r}$. Comme (cf. [11])

$$\mathrm{E}(S_n - np)^4 = n(p(1-p)^4 + (1-p)p^4) + 3n(n-1)p^2(1-p)^2 \leqslant np + 3n^2p^2,$$

une inégalité de type Tchébychev nous donne

$$\mathrm{P}\left(\left|w^n\left(\frac{k}{2^r}\right) - w^n\left(\frac{k-1}{2^r}\right)\right| \geqslant \frac{\epsilon}{r^2}\right) = \mathrm{P}\left(|S_n - np| \geqslant \frac{\epsilon\sqrt{n}}{r^2}\right) \leqslant$$

$$\leqslant \frac{(np + 3n^2p^2)r^8}{\epsilon^4 n^2} = \frac{r^8}{\epsilon^4 2^r n} + \frac{3r^8}{\epsilon^4 2^{2r}}.$$

Le second membre de (5) est par conséquent au plus égal à

$$\sum_{l=l}^{m}\left[\frac{r^8}{\epsilon^4 n} + \frac{3r^8}{\epsilon^4 2^r}\right] \leqslant c\left(\frac{m^9}{\epsilon^4 n} + \frac{l^8}{\epsilon^4 2^l}\right).$$

où c est une constante absolue $\left(\sum_{r=l}^{m} r^8 \sim m^9/9 \text{ lorsque } m \to \infty,\ \sum_{r=l}^{\infty} r^8 2^{-r} \sim 2l^8 2^{-l} \text{ lorsque } l \to \infty\right)$. En posant $m = 3\log_2 n$, on obtient

$$\limsup_{n\to\infty} \mathrm{P}(\omega_\Delta^{[m]} \geqslant \epsilon) \leqslant c\,\frac{l^8}{\epsilon^4 2^l}.$$

Cette expression peut être rendue aussi petite que l'on veut par un choix convenable de l (ou de Δ).

Estimons maintenant le deuxième terme de (4) qui est au plus égal à

$$2^m \mathrm{P}\left(\omega\left(\frac{k-1}{2^m}, \frac{k}{2^m}\right) \geqslant \epsilon\right). \tag{6}$$

L'événement de (6) exprime que si l'on fixe m, l'écart entre $n(F_n^*(u) - u)$ et $n(F_n^*(k/n^3) - k/n^3)$ sera supérieur à $\sqrt{n}$ sur l'intervalle $](k-1)/n^3, k/n^3[$ de largeur n^{-3}. Comme $\sqrt{n}\,\epsilon \geqslant 3$ pour n assez grand, il faut pour cela que l'intervalle $](k-1)/n^3, k/n^3[$ contienne au moins 2 éléments de l'échantillon X. Autrement dit, l'événement $\{s_n \geqslant 2\}$ doit se produire, si l'on se sert des notations de la série d'épreuves de Bernoulli, pour $p = n^{-3}$. Mais comme $1 = (1 - p + p)^n = (1-p)^n + np(1-p)^{n-1} + O(n^2p^2)$, il vient

$$\mathrm{P}(S_n \geqslant 2) = 1 - (1-p)^n - np(1-p)^{n-1} = O(n^2p^2).$$

Donc, (6) est au plus égal à $n^3O(n^{-4}) = O(n^{-4}) = o(1)$. ◄

Démonstration du théorème 1. Pour tout $x \in D(0, 1)$, posons

$$|x| = \sup_{0\leqslant t\leqslant 1}|x(t)|,\quad f_\epsilon^+(x) = \sup_{|y-x|\leqslant\epsilon} f(y),\quad f_\epsilon^-(x) = \inf_{|y-x|\leqslant\epsilon} f(y)$$

et désignons par x_Δ la ligne polygonale continue de nœuds $(k\Delta, x(k\Delta) = x_\Delta(k\Delta))$, $k = 0, \ldots$ $\ldots, 1/\Delta$, où Δ est une partie aliquote de 1. Remarquons que

$$|x - x_\Delta| \leqslant \omega_\Delta(x) \tag{7}$$

et que $f_\epsilon^{+}(x_\Delta)$ sont des fonctions continues du vecteur $(x(0), x(\Delta), x(2\Delta), \ldots, x(1))$. Le lemme 1 et le deuxième théorème de continuité nous donnent pour $n \to \infty$

$$f_\epsilon^{+}(w_\Delta^n) \Rightarrow f_\epsilon^{+}(w_\Delta^0). \tag{8}$$

Par ailleurs, la continuité de w^0 et de la fonctionnelle f entraînent

$$|w_\Delta^0 - w^0| \leqslant \omega_\Delta(w^0) \underset{P}{\to} 0 \text{ pour } \Delta \to 0, \tag{9}$$

$$f_\epsilon^{+}(w^0) \underset{P}{\to} f(w^0) \text{ pour } \epsilon \to 0. \tag{10}$$

De la définition de f_ϵ^- il découle que $f_\epsilon^-(y) \leqslant f(x)$ sur l'ensemble $|y - x| \leqslant \epsilon$. Donc

$$P(f(w^n) \leqslant t) \leqslant P(f_\epsilon^-(w_\Delta^n) \leqslant t, |w_\Delta^n - w^n| \leqslant \epsilon) + P(|w_\Delta^n - w^n| > \epsilon) \leqslant$$
$$\leqslant P(f^-(w_\Delta^n) \leqslant t) + P(\omega_\Delta(w^n) > \epsilon).$$

En passant à la limite pour $n \to \infty$ et en se servant de (8) et (9), on obtient

$$\limsup_{n\to\infty} P(f(w^n) \leqslant t) \leqslant P(f_\epsilon^-(w_\Delta^0) \leqslant t) + \limsup_{n\to\infty} P(\omega_\Delta(w^n) > \epsilon). \tag{11}$$

On trouve de façon analogue

$$P(f_\epsilon^-(w_\Delta^0) \leqslant t) \leqslant P(f_{2\epsilon}^-(w^0) \leqslant t) + P(\omega_\Delta(w^0) > \epsilon).$$

Portons maintenant la dernière expression dans (11) et passons à la limite lorsque $\Delta \to 0$. De (9) et du lemme 2, on déduit alors que

$$\limsup_{n\to\infty} P(f(w^n) \leqslant t) \leqslant P(f_{2\epsilon}^-(w^0) \leqslant t).$$

De là et de (10), il s'ensuit

$$\limsup_{n\to\infty} P(f(w^n) \leqslant t) \leqslant P(f(w^0) \leqslant t).$$

On établit de façon analogue l'inégalité contraire

$$\liminf_{n\to\infty} P(f(w^n) < t) \geqslant P(f(w^0) < t).$$

Ces inégalités expriment de toute évidence que $f(w^n) \Rightarrow f(w^0)$. ◄

Considérons encore un théorème limite fonctionnel pour des processus empiriques, qui rappelle beaucoup le théorème 1.

Supposons qu'en plus de l'échantillon X de taille n_1 nous est donné un échantillon Y de taille n_2 indépendant de X et distribué suivant la même loi uniforme sur $[0, 1]$. Pour la commodité on désignera ici les fonctions de répartition empiriques de X et Y respectivement par $F_X^*(t)$ et $F_Y^*(t)$. Posons

$$w_{X,Y}(t) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,(F_X^*(t) - F_Y^*(t)).$$

Théorème 2. *Si une fonctionnelle f satisfait les conditions du théorème* 1, *alors pour* $n_1 \to \infty$, $n_2 \to \infty$

$$f(w_{X,Y}) \Rightarrow f(w^0).$$

Démonstration. Prouvons ce théorème sous la condition simplificatrice que

$$a = \frac{n_2}{n_1 + n_2} \to a_0 \in [0, 1]$$

lorsque $n \to \infty$. On a

$$w_{X,Y}(t) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,[(F_X^*(t) - t) - (F_Y^*(t) - t)] = \sqrt{a}\,w_X(t) + \sqrt{1-a}\,w_Y(t), \tag{12}$$

où $w_X(t)$ et $w_Y(t)$ sont des processus empiriques correspondant aux échantillons X et Y.

Comme $\omega_\Delta(x + y) \leqslant \omega_\Delta(x) + \omega_\Delta(y)$, on déduit aussitôt de (12) et du lemme 2 l'analogue du lemme 2 pour le processus $w_{X,Y}(t)$: pour tout $\epsilon > 0$

$$\limsup_{n\to\infty} \mathbf{P}(\omega_\Delta(w_{X,Y}) > \epsilon) \to 0.$$

La convergence des distributions finidimensionnelles de $w_{X,Y}$ et de w° découle également de (12). En effet, désignons par $\mathbf{w}_{X,Y}$, $\mathbf{w}_X$, $\mathbf{w}_Y$ les vecteurs construits d'après les processus $w_{X,Y}(t)$, $w_X(t)$, $w_Y(t)$ exactement comme le vecteur $\mathbf{w}^n$ l'a été d'après $w^n(t)$. En s'appuyant alors sur l'indépendance de X et Y et sur la démonstration du lemme 1 on obtient

$$\mathbf{E}e^{i\mathbf{w}_{X,Y}\mathbf{u}^T} = \mathbf{E}e^{i\sqrt{a}\,\mathbf{w}_X\mathbf{u}^T}\mathbf{E}e^{i\sqrt{1-a}\,\mathbf{w}_Y\mathbf{u}^T} \to \exp\left\{-\frac{a_0 + (1 - a_0)}{2}\left(\sum_{j=0}^{m} u_j^2\Delta_j - U^2\right)\right\} =$$

$$= \exp\left\{-\frac{1}{2}\left(\sum_{j=0}^{m} u_j^2\Delta_j - U^2\right)\right\} = \mathbf{E}e^{i\mathbf{w}^\circ\mathbf{u}^T}.$$

Pour le reste la démonstration du théorème 2 est calquée sur celle du théorème 1. ◄

ANNEXE III

PROPRIÉTÉS DES ESPÉRANCES MATHÉMATIQUES CONDITIONNELLES

Dans le § 2.9 nous avons énuméré les principales propriétés de l'espérance mathématique conditionnelle. On produit plus bas les démonstrations de ces propriétés dans l'ordre de leur énumération dans le § 2.9.

1a. $\mathbf{E}(c\xi|\mathfrak{A}) = c\mathbf{E}(\xi|\mathfrak{A})$.

1b. $\mathbf{E}(\xi_1 + \xi_2|\mathfrak{A}) = \mathbf{E}(\xi_1|\mathfrak{A}) + \mathbf{E}(\xi_2|\mathfrak{A})$.

1c. *Si* $\xi_1 \leqslant \xi_2$ *p.s., alors* $\mathbf{E}(\xi_1|\mathfrak{A}) \leqslant \mathbf{E}(\xi_2|\mathfrak{A})$ *p.s.*

Pour établir la propriété 1a, il faut s'assurer en vertu de la définition 2.9.2 que

1) $c\mathbf{E}(\xi|\mathfrak{A})$ est une fonction $\mathfrak{A}$-mesurable,
2) $\mathbf{E}(c\mathbf{E}(\xi|\mathfrak{A}); A) = \mathbf{E}(c\xi; A)$ pour tout $A \in \mathfrak{A}$.

La première propriété est évidente. La deuxième découle des propriétés de linéarité de l'espérance mathématique ordinaire (ou d'une intégrale ordinaire) :

$$\mathbf{E}(c\mathbf{E}(\xi|\mathfrak{A}); A) = c\mathbf{E}(\mathbf{E}(\xi|\mathfrak{A}); A) = c\mathbf{E}(\xi; A) = \mathbf{E}(c\xi; A).$$

La propriété 1b s'obtient de la même façon.

Pour prouver la propriété 1c, on posera pour simplifier $\hat{\xi}_i = \mathbf{E}(\xi_i|\mathfrak{A})$. Alors, pour tout $A \in \mathfrak{A}$

$$\int_A \hat{\xi}_1 d\mathbf{P} = \mathbf{E}(\hat{\xi}_1; A) = \mathbf{E}(\xi_1; A) \leqslant \mathbf{E}(\xi_2; A) = \int_A \hat{\xi}_2 d\mathbf{P},$$

$$\int_A (\hat{\xi}_2 - \hat{\xi}_1) d\mathbf{P} \geqslant 0.$$

D'où il s'ensuit que $\hat{\xi}_2 - \hat{\xi}_1 \geqslant 0$ p.s.

2. *Inégalité de Tchébychev. Si* $\xi \geqslant 0$, $x \geqslant 0$, *on a*

$$\mathbf{P}(\xi \geqslant x|\mathfrak{A}) \leqslant \frac{\mathbf{E}(\xi|\mathfrak{A})}{x}.$$

Cette propriété découle de 1c puisque $\mathbf{P}(\xi \geqslant x|\mathfrak{A}) = \mathbf{E}(I_{\{\xi \geqslant x\}}|\mathfrak{A})$, où I_A est l'indicateur de l'événement A, et que $I_{\{\xi \geqslant x\}} \leqslant \xi/x$.

3. *Si* $\mathfrak{A}$ *et* $\sigma(\xi)$ *sont indépendantes*, $\mathbf{E}(\xi|\mathfrak{A}) = \mathbf{E}\xi$. Comme $\hat{\xi} = \mathbf{E}\xi$ est une fonction $\mathfrak{A}$-mesurable, il reste à prouver seulement la deuxième condition de la définition 2.9.2 : pour tout $A \in \mathfrak{A}$, on a

$$\mathbf{E}(\hat{\xi}; A) = \mathbf{E}(\xi; A).$$

La véracité de cette égalité découle de l'indépendance des variables aléatoires I_A et ξ et des relations

$$\mathbf{E}(\xi;A) = \mathbf{E}(\xi I_A) = \mathbf{E}\xi\cdot\mathbf{E}I_A = \hat{\xi}\mathbf{P}(A) = \mathbf{E}(\hat{\xi};A).$$

4. *Théorème de convergence monotone. Si* $0 \leqslant \xi_n \uparrow \xi$ *p.s., alors* $\mathbf{E}(\xi_n|\mathfrak{A}) \uparrow \mathbf{E}(\xi|\mathfrak{A})$ *p.s.* En effet, la relation $\xi_{n+1} \geqslant \xi_n$ p.s. entraîne $\hat{\xi}_{n+1} \geqslant \hat{\xi}_n$ p.s., où $\hat{\xi}_n = \mathbf{E}(\xi_n|\mathfrak{A})$. Il existe donc une variable aléatoire $\hat{\xi}$ $\mathfrak{A}$-mesurable telle que $\hat{\xi}_n \uparrow \hat{\xi}$ p.s. Le théorème ordinaire de convergence monotone nous dit que pour tout $A \in \mathfrak{A}$

$$\int_A \hat{\xi}_n d\mathbf{P} \to \int_A \hat{\xi} d\mathbf{P}, \quad \int_A \xi_n d\mathbf{P} \to \int_A \xi d\mathbf{P}.$$

Les premiers membres de ces relations étant confondus, il en sera de même des seconds. Ce qui exprime que $\hat{\xi} = \mathbf{E}(\xi|\mathfrak{A})$.

5. *Si* η *est réelle et* $\mathfrak{A}$*-mesurable, alors*

$$\mathbf{E}(\eta\xi|\mathfrak{A}) = \eta\mathbf{E}(\xi|\mathfrak{A}). \tag{1}$$

Si $\eta = I_B$ (l'indicateur de $B \in \mathfrak{A}$), cette proposition est vraie puisque pour tout $A \in \mathfrak{A}$

$$\int_A \mathbf{E}(I_B\xi|\mathfrak{A})d\mathbf{P} = \int_A I_B\xi d\mathbf{P} = \int_{AB} \xi d\mathbf{P} = \int_{AB} \mathbf{E}(\xi|\mathfrak{A})d\mathbf{P} = \int_A I_B\mathbf{E}(\xi|\mathfrak{A})d\mathbf{P}.$$

De là et de la linéarité de l'espérance mathématique conditionnelle il résulte que cette proposition est valable aussi pour toute fonction simple η.

Si $\xi \geqslant 0$ et $\eta \geqslant 0$, en considérant une suite de fonctions simples $0 \leqslant \eta_n \uparrow \eta$ et en appliquant le théorème de convergence monotone à l'égalité

$$\mathbf{E}(\eta_n\xi|\mathfrak{A}) = \eta_n\mathbf{E}(\xi|\mathfrak{A}),$$

on obtient (1). Le passage à des ξ et η arbitraires s'effectue comme d'habitude en considérant les parties positives et négatives des variables aléatoires ξ et η. Ceci étant, pour que les différences et sommes obtenues aient un sens, il faut exiger l'existence de $\mathbf{E}|\xi| < \infty$ et $\mathbf{E}|\xi\eta| < \infty$.

6. *L'inégalité de Cauchy-Bouniakovski*

$$\mathbf{E}(\xi_1\xi_2|\mathfrak{A}) \leqslant [\mathbf{E}(\xi_1^2|\mathfrak{A})\mathbf{E}(\xi_2^2|\mathfrak{A})]^{1/2}$$

se prouve exactement comme pour les espérances mathématiques ordinaires (cf. par exemple [11]), puisque la démonstration n'utilise aucune propriété des espérances mathématiques hormis la linéarité.

L'inégalité de Jensen

$$g(\mathbf{E}(\xi|\mathfrak{A})) \leqslant \mathbf{E}(g(\xi)|\mathfrak{A}) \tag{2}$$

pour une fonction g convexe vers le bas découle des relations suivantes (comparer avec [11]). La fonction $g(x)$ étant convexe, pour tout y il existe un nombre $g_1(y)$ tel que

$$g(x) \leqslant g(y) + (x - y)g_1(y).$$

Posons ici $x = \xi$, $y = \hat{\xi} = \mathbf{E}(\xi|\mathfrak{A})$ et prenons l'espérance mathématique conditionnelle des deux parties de cette inégalité. La relation annoncée résulte de ce que

$$\mathbf{E}[(\xi - \hat{\xi})g_1(\hat{\xi})|\mathfrak{A}] = g_1(\hat{\xi})\mathbf{E}[\xi - \hat{\xi}|\mathfrak{A}] = 0,$$

en vertu de la propriété 5.

7. *La formule des probabilités totales* découle de la propriété 8 si pour $\mathfrak{A}$ on prend une tribu triviale.

8. *Si* $\mathfrak{A} \subset \mathfrak{A}_1 \subset \mathfrak{F}$, *on a la formule de « moyennisation successive »*

$$\mathsf{E}(\xi|\mathfrak{A}) = \mathsf{E}(\mathsf{E}(\xi|\mathfrak{A}_1)|\mathfrak{A}).$$

En effet, pout tout $A \in \mathfrak{A}$ et du fait que $A \in \mathfrak{A}_1$ il vient

$$\int_A \mathsf{E}(\mathsf{E}(\xi|\mathfrak{A}_1)|\mathfrak{A})d\mathbf{P} = \int_A \mathsf{E}(\xi|\mathfrak{A}_1)d\mathbf{P} = \int_A \xi d\mathbf{P} = \int_A \mathsf{E}(\xi|\mathfrak{A})d\mathbf{P}.$$

Signalons en conclusion que la propriété 5 admet la généralisation suivante sous des conditions larges.

5A. *Si* η *est* $\mathfrak{A}$*-mesurable et* $\varphi(\omega, \eta)$ *est une fonction mesurable des variables* $\omega \in \Omega$ *et* $\eta \in$ $\in R^k$, *alors*

$$\mathsf{E}(\varphi(\omega, \eta)|\mathfrak{A}) = \psi(\omega, \eta), \text{ où } \psi(\omega, y) = \mathsf{E}(\varphi(\omega, y)|\mathfrak{A}). \tag{3}$$

On prouvera cette propriété sous l'hypothèse qu'il existe une suite de fonctions simples η_n, telle que $\varphi(\omega, \eta_n) \uparrow \varphi(\omega, \eta)$, $\psi(\omega, \eta_n) \uparrow \psi(\omega, \eta)$ p.s. En effet, supposons que $\eta_n = y_k$ pour $\omega \in A_k \subset \mathfrak{A}$. Alors

$$\varphi(\omega, \eta_n) = \sum_k \varphi(\omega, y_k) I_{A_k}.$$

De là on déduit que (3) est réalisée pour les fonctions η_n en vertu de la propriété 5. Reste à appliquer le théorème de convergence monotone (propriété 4) à l'égalité

$$\mathsf{E}(\varphi(\omega, \eta_n)|\mathfrak{A}) = \psi(\omega, \eta_n).$$

ANNEXE IV

THÉORÈME DE FACTORISATION DE NEYMAN-FISHER

On prouve ici le théorème 2.12.1.

Pour alléger les notations on admettra sans perte de généralité que $n = 1$ (en effet l'échantillon X peut être multidimensionnel). D'autre part, puisque nous avons convenu que l'espace $(\mathscr{X}, \mathfrak{B})$ est l'espace des échantillons, on écrira $\mathbf{P}_\theta(B)$ au lieu de $\mathbf{P}_\theta(X \in B)$. La dimension de la statistique S sera désignée par l.

THÉORÈME 1. *Soit remplie la condition (A_μ). Une statistique S est exhaustive si et seulement s'il existe une fonction $\psi(\theta, s)$ positive mesurable par rapport à $s \in R^l$ et une fonction $h(x)$ positive mesurable par rapport à $x \in \mathscr{X}$, telles que*

$$f_\theta(x) = \frac{d\mathbf{P}_\theta}{d\mu}(x) = \psi(\theta, S(x))\, h(x), \quad |\mu|\text{-p.p.} \tag{1}$$

Démontrons préalablement deux propositions auxiliaires. Introduisons la

CONDITION (D). *La famille $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ satisfait la condition (A_λ)* (c'est-à-dire est dominée par la mesure λ), *où la mesure de probabilité λ est de la forme*

$$\lambda = \sum_i c_i \mathbf{P}_{\theta_i},\ \theta_i \in \Theta,\ c_i > 0,\ \sum_i c_i = 1.$$

THÉORÈME 2. *La condition (A_μ) est nécessaire et suffisante à la réalisation de la condition (D).*

DÉMONSTRATION. La nécessité est évidente. Prouvons la suffisance. Sans perdre en généralité, on peut admettre que μ est une mesure de probabilité. En effet, on peut toujours remplacer la mesure μ par la mesure

$$\mu^*(A) = \sum_j \frac{\mu(AB_j)}{2^j \mu(B_j)},$$

où $\{B_j\}$ est une partition de l'espace $\mathscr{X}$, telle que $\mu(B_j) < \infty$, $j = 1, 2, \ldots$

Soit $\tilde{\mathscr{P}}$ la classe de toutes les mesures de probabilité de la forme $\mathbf{P} = \sum c_i \mathbf{P}_{\theta_i}$, $\theta_i \in \Theta$, $c_i > 0$, $\sum c_i = 1$. Il est évident que $\tilde{\mathscr{P}} \supset \mathscr{P}$ et vérifie aussi la condition (A_μ).

Désignons $p = d\mathbf{P}/d\mu$ et considérons la classe $\mathfrak{C}$ des ensembles $C \in \mathfrak{B}$ pour lesquels il existe un $P \in \tilde{\mathscr{P}}$ tel que $p(x) > 0$ p.s. sur C, $\mathbf{P}(C) > 0$. Soit $C_1, C_2, \ldots$ une suite d'ensembles de $\mathfrak{C}$ telle que

$$\mu(C_i) \to \sup_{C\in\mathfrak{C}} \mu(C).$$

Comme $C_i \in \mathfrak{C}$, il existe un $\mathbf{P}^{(i)} \in \mathscr{P}$ tel que $p^{(i)} = \dfrac{d\mathbf{P}^{(i)}}{d\mu} > 0$ p.s. sur C_i. Posons

$$C_0 = \cup C_i, \quad \mathbf{P}^{(0)} = \sum_i c_i \mathbf{P}^{(i)}, \quad p^{(0)} = \sum_i c_i p^{(i)}$$

pour certains $c_i > 0$, $\sum c_i = 1$. Il est évident que $p^{(0)} > 0$ sur C_0 et par suite $C_0 \in \mathfrak{C}$.

On prouvera la proposition du théorème lorsqu'on aura établi que $\mathbf{P}^{(0)}(A) = 0$ entraîne $\mathbf{P}(A) = 0$ pour tous les $\mathbf{P} \in \mathscr{P}$. Ceci exprimera que $\mathbf{P}_\theta$ est absolument continue par rapport à $\lambda = \mathbf{P}^{(0)}$ et que la condition (D) est réalisée.

Supposons donc que $\mathbf{P}^{(0)}(A) = 0$ et soit $\mathbf{P}$ un autre élément quelconque de $\mathscr{P}$. Désignons $C = \{x : p(x) > 0\}$. La proposition annoncée résultera des trois relations suivantes :

$$\mathbf{P}(AC_0) = 0, \quad \mathbf{P}(A\overline{C}_0\overline{C}) = 0, \quad \mathbf{P}(A\overline{C}_0 C) = 0,$$

où $\overline{B}$ est le complémentaire de B. La première de ces relations découle du fait que $\mathbf{P}^{(0)}(AC_0) = 0$, $p^{(0)}(x) > 0$ sur C_0 et donc $\mu(AC_0) = 0$. La deuxième, du fait que $p(x) = 0$ sur $\overline{C}$. Prouvons la troisième par l'absurde. En admettant que $R = A\overline{C}_0 C$ on trouve $\mu(R) > 0$, $\mu(C_0 \cup R) - \mu(C_0) > 0$. Ce qui contredit l'égalité

$$\mu(C_0) = \sup_{C \in \mathfrak{C}} \mu(C),$$

puisque $C_0 \in \mathfrak{C}$, $R \in \mathfrak{C}$, $C_0 \cup R \in \mathfrak{C}$. ◄

Ainsi, nous avons établi que si la condition (A_μ) est réalisée, il existe une mesure λ pour laquelle est remplie la condition (D).

THÉORÈME 3. *Une statistique S est exhaustive si et seulement s'il existe une fonction $g_\theta(s)$ mesurable telle que*

$$\frac{d\mathbf{P}_\theta}{d\lambda}(x) = g_\theta(S(x)) \quad [\lambda]\text{-p.p.} \tag{2}$$

DÉMONSTRATION. Pour tout $B \subset R^l$ mesurable, posons $S^{-1}(B) = \{x \in \mathscr{X} : S(x) \in B\} \in \mathfrak{B}_x$ et considérons la distribution $\mathbf{G}_\theta$ de la statistique S induite sur R^l par la distribution $\mathbf{P}_\theta$:

$$\mathbf{G}_\theta(B) = \int_{S^{-1}(B)} \mathbf{P}_\theta(dx) = \int_{S^{-1}(B)} \frac{d\mathbf{P}_\theta}{d\lambda}(x)\lambda(dx).$$

Considérons aussi la distribution

$$\nu(B) = \int_{S^{-1}(B)} \lambda(dx).$$

Il est clair que $\mathbf{G}_\theta$ est absolument continue par rapport à ν, puisque $\nu(B) = 0$ entraîne $\mathbf{G}_\theta(B) = 0$. Il existe donc une densité $g_\theta(s)$ mesurable par rapport à s telle que

$$\mathbf{G}_\theta(B) = \int_B g_\theta(s)\nu(ds).$$

Supposons maintenant que S est une statistique exhaustive, donc qu'il existe une distribution conditionnelle $P(A|s) = \mathbf{P}_\theta(A|S(x) = s)$ indépendante de θ. Par définition de la distribution conditionnelle, pour tout $A_0 \in \sigma(S)$, on a

$$\int_{A_0} P(A|S(x))\mathbf{P}_\theta(dx) = \mathbf{P}_\theta(A \cap A_0).$$

De là il s'ensuit également que

$$\int_{A_0} P(A|S(x))\lambda(dx) = \lambda(A \cap A_0).$$

Ceci exprime que $P(A|S)$ est simultanément une probabilité conditionnelle par rapport à λ. On désignera cette probabilité comme l'espérance mathématique conditionnelle $\mathbf{E}_\lambda(I_A|S)$ de l'indicateur I_A.

Pour $A_0 = R^l$ on déduit de (1), en vertu des propriétés de l'espérance mathématique conditionnelle,

$$\mathbf{P}_\theta(A) = \int P(A|S(x))\mathbf{P}_\theta(dx) = \mathbf{E}_\theta P(A|S(X)) = \int P(A|s)\mathbf{G}_\theta(ds) = \int P(A|S)g_\theta(s)\nu(ds) =$$

$$= \int P(A|S(x))g_\theta(S(x))\lambda(dx) = \int \mathbf{E}_\lambda(I_A|S(x))g_\theta(S(x))\lambda(dx) =$$

$$= \int \mathbf{E}_\lambda(I_A g_\theta(S(x))|S(x))\lambda(dx) = \int I_A g_\theta(S(x))\lambda(dx) = \int_A g_\theta(S(x))\lambda(dx).$$

Ce qui, de toute évidence, équivaut à (2).

Supposons maintenant qu'est remplie (2). On prouvera que l'espérance mathématique conditionnelle $\mathbf{E}_\lambda(I_A|S)$ associée à la distribution λ (cette espérance ne dépend pas de θ) est en même temps l'espérance mathématique conditionnelle $\mathbf{P}_\theta(A|S)$ pour tous les $\mathbf{P}_\theta \in \mathscr{P}$.

Fixons A et θ et introduisons une mesure γ sur $\mathfrak{B}$ à l'aide de l'égalité

$$\gamma(C) = \mathbf{P}_\theta(AC), \ C \in \mathfrak{B},$$

de sorte que $d\gamma/d\mathbf{P}_\theta = I_A$, $d\gamma/d\lambda = I_A g_\theta(S(x))$.

Pour tout $C \in \sigma(S)$, on a

$$\gamma(C) = \int_C I_A \mathbf{P}_\theta(dx) = \mathbf{E}_\theta I_A I_C = \mathbf{E}_\theta I_C \mathbf{E}_\theta(I_A|S) = \int_C \mathbf{E}_\theta(I_A|S)\mathbf{P}_\theta(dx). \quad (3)$$

Si donc l'on traite γ, $\mathbf{P}_\theta$ et λ comme des distributions sur $\sigma(S)$, on obtient

$$\frac{d\gamma}{d\mathbf{P}_\theta} = \mathbf{E}_\theta(I_A|S),$$

$$\frac{d\gamma}{d\lambda} = \mathbf{E}_\theta(I_A|S)\frac{d\mathbf{P}_\theta}{d\lambda} = \mathbf{E}_\theta(I_A|S)g_\theta(S).$$

Par analogie à (3) on a sur $\sigma(S)$

$$\frac{d\gamma}{d\lambda} = \mathbf{E}_\lambda(I_A g_\theta(S)|S) = g_\theta(S)\mathbf{E}_\lambda(I_A|S).$$

D'où il s'ensuit que λ-p.s. (ici et plus bas par λ et $\mathbf{P}_\theta$ on comprendra des distributions sur $\sigma(S)$)

$$\mathbf{E}_\theta(I_A|S)g_\theta(S) = \mathbf{E}_\lambda(I_A|S)g_\theta(S). \quad (4)$$

Utilisons maintenant la propriété (D) qui dit que si (4) est réalisée λ-p.s., elle le sera $\mathbf{P}_\theta$-p.s. Par ailleurs, on a $\mathbf{P}_\theta$-p.s.

$$g_\theta(S(x)) = \frac{d\mathbf{P}_\theta}{d\lambda}(x) \neq 0.$$

Donc

$$\mathbf{P}_\theta(A|S) = \mathbf{E}_\theta(I_A|S) = \mathbf{E}_\lambda(I_A|S), \quad \mathbf{P}_\theta\text{-p.s.}$$

Ce qui exprime que la quantité $\mathbf{E}_\lambda(I_A|S)$ qui est indépendante de θ peut être prise pour probabilité conditionnelle $\mathbf{P}_\theta(A|S)$. ◄

DÉMONSTRATION du théorème 1. Si S est une statistique exhaustive, la relation (1) découle du théorème 3, puisque

$$f_\theta(x) = \frac{d\mathbf{P}_\theta}{d\mu} = g_\theta(S(x)) \frac{d\lambda}{d\mu}(x),$$

où il faut poser $g_\theta(s) = \psi(\theta, s)$, $\frac{d\lambda}{d\mu}(x) = h(x)$. Réciproquement si (1) a lieu, alors

$$\frac{d\lambda}{d\mu} = \sum c_i \frac{d\mathbf{P}_{\theta_i}}{d\mu} = \sum c_i \psi(\theta_i, S(x))h(x) = r(S(x))h(x).$$

Donc, si $r(S(x)) > 0$, alors

$$\frac{d\mathbf{P}_\theta}{d\lambda} = \frac{d\mathbf{P}_\theta}{d\mu} \cdot \frac{d\mu}{d\lambda} = \frac{\psi(\theta, S(x))}{r(Sx))}.$$

Si $r(S(x)) = 0$, on peut définir $\frac{d\mathbf{P}_\theta}{d\lambda}(x)$ de façon arbitraire, puisque λ est une mesure et par suite $\mathbf{P}_\theta$, la mesure de l'ensemble de tels points x, est nulle. En posant $g_\theta(s) = \psi(\theta, s)/r(s)$ et en appliquant le théorème 3, on trouve que S est une statistique exhaustive. ◄

ANNEXE V

LOI DES GRANDS NOMBRES ET THÉORÈME LIMITE CENTRAL. VARIANTES UNIFORMES

1. Loi des grands nombres dans le schéma des séries. Considérons une suite $\{\xi_{k.n}\}_{k=1}^{n}$, $n = 1, 2, \ldots$, de vecteurs indépendants équidistribués dans un schéma de séries (la distribution de $\xi_{k.n}$ dépend de n) et supposons que $\mathbf{E}\xi_{k.n} = 0$.

Désignons $\zeta_n = \sum_{k=1}^{n} \xi_{k.n}$.

THÉORÈME 1. *Supposons que*

$$n\mathbf{E}|\xi_{k.n}| = a_n < a < \infty,$$
$$n\mathbf{E}(|\xi_{k.n}|;\ |\xi_{k.n}| > \tau) \to 0 \tag{1}$$

lorsque $n \to \infty$ *pour tout* $\tau > 0$. *Alors, pour tout* $\epsilon > 0$

$$\mathbf{P}(|\zeta_n| > \epsilon) \to 0.$$

DÉMONSTRATION: Considérons les variables aléatoires $\xi'_{k.n}$ obtenues par troncature de $\xi_{k.n}$ au niveau τ :

$$\xi'_{k.n} = \begin{cases} \xi_{k.n} & \text{si} \quad |\xi_{k.n}| \leqslant \tau, \\ 0 & \text{si} \quad |\xi_{k.n}| > \tau. \end{cases}$$

En vertu de la condition (1)

$$\mathbf{P}(\xi'_{1.n} \neq \xi_{1.n}) = \mathbf{P}(|\xi_{1.n}) > \tau) \leqslant \frac{1}{\tau}\mathbf{E}(|\xi_{1.n}|;\ |\xi_{1.n}| > \tau) = o(1/n);\ \mathbf{E}\xi'_{1.n} = o(1/n),$$
$$\mathbf{E}(\xi'_{1.n})^2 = \mathbf{E}(\xi^2_{1.n};\ |\xi_{1.n}| \leqslant \tau) \leqslant \tau\mathbf{E}(|\xi_{1.n}|;\ |\xi_{1.n}| \leqslant \tau) = \tau(a_n/n - \mathbf{E}(|\xi_{1.n}|;\ |\xi_{1.n}| > \tau)).$$

Donc, pour tout $\varepsilon > 0$ et n assez grand

$$\mathbf{E}(\xi'_{1.n})^2 \leqslant 2a\tau/n, \quad \mathbf{V}\xi'_{1.n} \leqslant 2a\tau/n, \quad n\mathbf{E}\xi'_{1.n} < \epsilon/2.$$

Posons $\zeta'_n = \sum_{j=1}^{n} \xi'_{j.n}$. Pour les n assez grands, on a alors

$$\mathbf{P}(|\zeta_n| > \epsilon) \leqslant \mathbf{P}\Big(\bigcup_{j=1}^{n} \{\xi'_{j.n} \neq \xi_{j.n}\}\Big) + \mathbf{P}(|\zeta'_n| > \epsilon).$$

Le premier terme est au plus égal à $n\mathbf{P}(\xi'_{1.n} \neq \xi_{1.n}) = o(1)$, le second, à

$$\mathbf{P}(|\zeta'_n - \mathbf{E}\zeta'_n| > \epsilon/2) \leqslant 4\mathbf{V}\zeta'_n/\epsilon^2 \leqslant 8a\tau/\epsilon^2.$$

Puisque τ est arbitraire, la valeur obtenue peut être rendue aussi petite que l'on veut quel que soit $\varepsilon > 0$. En choisissant maintenant n assez grand, on peut rendre la probabilité $\mathbf{P}(|\zeta_n| > \varepsilon)$ arbitrairement petite. ◄

2. Théorème limite central dans un schéma de séries. On admettra que

$$\mathbf{E}\xi_{j,n} = 0, \ \mathbf{E}|\xi_{j,n}|^2 < \infty.$$

Désignons $\sigma_n^2 = n\mathbf{E}\xi_{1,n}^T\xi_{1,n}, \ \zeta_n = \sum_{j=1}^n \xi_{j,n}$.

Théorème 2. *Supposons remplies les conditions de Lindeberg*

$$n\mathbf{E}(|\xi_{1,n}|^2; \ |\xi_{1,n}| > \tau) \to 0, \quad n \to \infty,$$

pour tout $\tau > 0$. *Si* $\sigma_n^2 \to \sigma^2$, *alors*

$$\zeta_n \in \Phi_{0,\sigma^2}.$$

Corollaire 1 (théorème limite central ordinaire). *Si* $\xi_1, \xi_2, \ldots$ *est une suite de vecteurs indépendants équidistribués,* $\mathbf{E}\xi_k = 0$, $\sigma^2 = \mathbf{E}\xi_k^T\xi_k < \infty$, $s_n = \sum_{k=1}^n \xi_k$, *alors pour* $n \to \infty$

$$\frac{s_n}{\sqrt{n}} \in \Phi_{0,\sigma^2}.$$

Cette proposition découle du théorème 2, puisque les variables aléatoires $\xi_{k,n} = \xi_k/\sqrt{n}$ vérifient les conditions dudit théorème.

Démonstration du théorème 2. Considérons les fonctions caractéristiques

$$\psi_n(t) = \mathbf{E}e^{i(t,\xi_{1,n})}, \quad \varphi_n(t) = \mathbf{E}e^{i(t,\zeta_n)} = \psi_n^n(t).$$

Pour prouver ce théorème, il faut s'assurer que pour tout t

$$\varphi_n(t) \to \exp\left\{-\frac{1}{2}t\sigma^2 t^T\right\}$$

lorsque $n \to \infty$.

Utilisons la version du théorème 1 établie pour le cas scalaire dans [11]. Les fonctions $\psi_n(t)$ et $\varphi_n(t)$ peuvent être traitées comme les fonctions caractéristiques

$$\psi_n^{(\omega)}(v) = \mathbf{E}e^{iv\xi_{1,n}^\omega} \quad \text{et} \quad \varphi_n^{(\omega)}(v) = \mathbf{E}e^{iv\zeta_n^\omega}$$

des variables aléatoires $\xi_{1,n}^\omega = (\xi_{1,n}, \omega)$, $\zeta_n^\omega = (\zeta_n, \omega)$, où $\omega = t/|t|$, $v = |t|$. Montrons que les variables aléatoires scalaires $\xi_{k,n}^\omega$ satisfont les conditions du théorème 1 pour le cas scalaire. Il est évident que

$$\mathbf{E}\xi_{k,n}^\omega = 0, \quad n\mathbf{E}(\xi_{1,n}^\omega)^2 = n\mathbf{E}(\xi_{1,n}, \omega)^2 = \omega\sigma_n^2\omega^T \to \omega\sigma^2\omega^T.$$

La réalisation de la condition de Lindeberg résulte de l'inégalité évidente

$$n\mathbf{E}((\xi_{1,n}, \omega)^2; \ |(\xi_{1,n}, \omega)| > \tau) \leqslant n\mathbf{E}(|\xi_{1,n}|^2; \ |\xi_{1,n}| > \tau).$$

Donc, pour tous v et ω (c'est-à-dire pour tout t)

$$\varphi_n(t) = \mathbf{E}e^{iv\zeta_n^\omega} \to \exp\left\{-\frac{1}{2}v^2\omega\sigma^2\omega^T\right\} = \exp\left\{-\frac{1}{2}t\sigma^2 t^T\right\}. \ ◄$$

3. Théorèmes limites uniformes pour les sommes de variables aléatoires dépendant d'un paramètre. On démontre ici les théorèmes 29.1 et 29.2.

Soient $X \in \mathbf{P}_\theta$ et $a(x, \theta)$ une fonction mesurable de $\mathscr{X} \times \Theta$ dans R^l donnée,

$$s_n(\theta) = \sum_{j=1}^{n} a(x_j, \theta).$$

On dira que l'intégrale $a(\theta) = \int a(x, \theta)\, \mathbf{P}_\theta(dx)$ converge uniformément en θ dans un domaine $\Theta_0 \subset \Theta$ si

$$\sup_{\theta \in \Theta_0} \int_{|a(x,\theta)| > N} |a(x, \theta)| \mathbf{P}_\theta(dx) \to 0$$

lorsque $N \to \infty$.

THÉORÈME 3 (loi uniforme des grands nombres). *Si l'intégrale* $a(\theta) = \int a(x, \theta)\mathbf{P}_\theta(dx)$ *converge uniformément en* θ *dans un domaine* $\Theta_0 \subset \Theta$, *alors*

$$\zeta_n(\theta) = \frac{s_n(\theta)}{n} - a(\theta) \underset{P_\theta}{\to} 0 \qquad (2)$$

uniformément en $\theta \in \Theta_0$.

DÉMONSTRATION. Supposons que (2) n'a pas lieu. Il existe alors $\epsilon > 0$, $\delta > 0$ et une suite $\theta_n \in \Theta_0$ tels que

$$\mathbf{P}_{\theta_n}\left(\left|\frac{\zeta_n(\theta_n)}{n}\right| > \epsilon\right) > \delta \qquad (3)$$

pour tous les n.

Considérons les variables aléatoires

$$\xi_{j,n} = \frac{a(x_j, \theta_n) - a(\theta_n)}{n}.$$

Il est aisé de voir qu'elles satisfont les conditions du théorème 1. En effet, posons $A_n = \{x: |a(x, \theta_n) - a(\theta_n)| > \tau n\}$. Alors

$$n\mathbf{E}_{\theta_n}|\xi_{j,n}| \leqslant 2a = 2 \sup_{\theta \in \Theta_0} \int |a(x, \theta)| \mathbf{P}_\theta(dx) < \infty,$$

$$n\mathbf{E}_{\theta_n}(|\xi_{j,n}|; |\xi_{j,n}| > \tau) = \int_{A_n} |a(x, \theta_n) - a(\theta_n)| \mathbf{P}_{\theta_n}(dx) \to 0.$$

La dernière relation résulte de la convergence uniforme de l'intégrale $a(\theta)$ et de l'inégalité de Tchébychev

$$\mathbf{P}_{\theta_n}(A_n) \leqslant \frac{\mathbf{E}_{\theta_n}|\xi_{1,n}|}{\tau} \leqslant \frac{2a}{\tau n} \to 0.$$

Ce qui vient d'être dit exprime que la suite $\{\xi_{j,n}\}$ vérifie la loi des grands nombres

$$\mathbf{P}_{\theta_n}\left(\left|\sum_{j=1}^{n} \xi_{j,n}\right| > \epsilon\right) \to 0$$

pour tout $\epsilon > 0$. Ceci contredit (3) et prouve le théorème. ◄

Passons au *théorème limite central.* Supposons que $\mathbf{E}_\theta a(x_1, \theta) = 0$.

Posons $\sigma^2(\theta) = \|\sigma_{ij}(\theta)\| = \mathbf{E}_\theta a^T(x_1, \theta)a(x_1, \theta)$ et désignons par $a_j(x, \theta), j = 1, \ldots, l$, les coordonnées des vecteurs $a(x, \theta)$.

THÉORÈME 4 (théorème limite central uniforme). *Supposons que les intégrales* $\sigma_{jj}(\theta) = \mathbf{E}_\theta a_j^2(x_1, \theta)$ *convergent uniformément dans* $\Theta_0 \subset \Theta$, *c'est-à-dire que*

$$\sup_{\Theta_0} \sigma_{jj}(\theta) < \infty,$$

$$\sup_{\Theta_0} \mathbf{E}_0(a_j^2(x_1, \theta); |a_j(x_1, \theta)| > N) \to 0 \tag{4}$$

lorsque $N \to \infty$. *Alors*

$$\frac{s_n(\theta)}{\sqrt{n}} \Subset \Phi_{0,\sigma^2(\theta)} \tag{5}$$

lorsque $n \to \infty$ *uniformément en* $\theta \in \Theta_0$.

DÉMONSTRATION. La non-réalisation de (5) exprime qu'il existe une suite $\theta_n \in \Theta_0$ pour laquelle les sommes des variables aléatoires $\xi_{j,n} = a(x_j, \theta_n)/\sqrt{n}$ ne convergeront pas en loi vers $\Phi_{0,\sigma^2(\theta_n)}$.

L'adhérence de $\{\sigma^2(\theta), \theta \in \Theta_0\}$ étant compacte, on peut admettre que la suite $\{\theta_n\}$ est choisie de telle sorte que pour une matrice σ^2 l'on ait

$$\sigma^2(\theta_n) = n\mathbf{E}_{\theta_n}\xi_{1,n}^T\xi_{1,n} \to \sigma^2. \tag{6}$$

La condition de non-réalisation de (5) exprime alors que $\sum_{j=1}^{n} \xi_{j,n}$ ne convergera pas en loi vers Φ_{0,σ^2}. Or ceci est impossible en vertu du théorème 2, puisque $\xi_{j,n}$ satisfont les conditions de ce théorème. En effet, en vertu de (6) il suffit de vérifier la condition de Lindeberg. Pour les ensembles $A_{i,n} = \{|a_i(x_1, \theta_n)| > \tau\sqrt{n}/l\}$

$$\sup_{\theta \in \Theta_0} \mathbf{P}_\theta(A_{i,n}) \leqslant \sup_{\theta \in \Theta_0} \frac{l\sigma_{ii}(\theta)}{n\tau^2} \to 0$$

lorsque $n \to \infty$. En utilisant le fait que $\{|\xi_{1,n}| > \tau\} \subset \bigcup_{i=1}^{l} A_{i,n}$, on trouve

$$n\mathbf{E}_{\theta_n}(|\xi_{1,n}|^2; |\xi_{1,n}| > \tau) \leqslant \sum_{i,k=1}^{l} \mathbf{E}_{\theta_n}(a_i^2(x_1, \theta_n); A_{k,n}). \tag{7}$$

Ici $\mathbf{E}_{\theta_n}(a_i^2(x_1, \theta_n); A_{i,n}) \to 0$ en vertu de la convergence uniforme de l'intégrale $\sigma_{ii}(\theta)$. Si $i \neq k$, en posant $B_{i,N} = \{|a_i(x_1, \theta_n)| > N\}$, on obtient

$$\mathbf{E}_{\theta_n}(a^2; A_{k,n}) = \mathbf{E}_{\theta_n}(a^2; A_{k,n}B_{i,N}) + \mathbf{E}_{\theta_n}(a^2; A_{k,n}\overline{B}_{i,N}).$$

Pour $\varepsilon > 0$ donné, on peut choisir N de telle sorte que le premier terme soit strictement inférieur à ε en vertu de (4). Le deuxième terme est au plus égal à $N^2\mathbf{P}_{\theta_n}(A_{k,n}) \to 0$ lorsque $n \to \infty$. Ce qui exprime que (7) tend vers 0 lorsque $n \to \infty$. ◄

ANNEXE VI

QUELQUES PROPOSITIONS RELATIVES AUX INTÉGRALES DÉPENDANT D'UN PARAMÈTRE

1. Théorèmes de convergence d'intégrales dépendant d'un paramètre. Soit $\{\psi(t, y)\}$ une famille de fonctions mesurables définies sur un espace mesurable $(\mathscr{Y}, \mathfrak{B}_y)$ muni d'une mesure ν. On s'occupera des conditions pour lesquelles

$$\int \psi(t, y)\nu(dy) \to \int \psi(\theta, y)\nu(dy) \quad \text{lorsque } t \to \theta. \tag{1}$$

Soit $\{A(t) = A(t, \theta), t \in \Theta\}$ une famille d'ensembles de $\mathfrak{B}_\psi$. Désignons par $I_{A(t)}(x)$ l'indicateur de $A(t)$ et par $\overline{A}(t)$ le complémentaire de $A(t)$.

La proposition suivante est une généralisation d'un théorème classique de Lebesgue.

THÉORÈME 1. *Supposons qu'une famille $\{A(t)\}$ est telle que*

1) $\psi(t, y)I_{A(t)}(y) \to \psi(\theta, y)$ *lorsque* $t \to \theta$ *pour* $[\nu]$*-presque toutes les valeurs* y *telles que* $\psi(\theta, y) \neq 0$.

2) $\sup_t|\psi(t, y)I_{A(t)}(y)| \leqslant \psi(y)$, *où* ψ *est une fonction intégrable* :

$$\int \psi(y)\nu(dy) < \infty.$$

Une condition nécessaire et suffisante pour que (1) *ait lieu est que*

$$\int \psi(t, y)I_{\overline{A}(t)}(y)\nu(dy) \to 0 \quad \textit{lorsque } t \to \theta. \tag{2}$$

DÉMONSTRATION. Le théorème de Lebesgue nous dit que

$$\int \psi(t, y)I_{A(t)}(y)\nu(dy) \to \int \psi(\theta, y)\nu(dy).$$

Comme

$$\int \psi = \int \psi I_A + \int \psi I_{\overline{A}},$$

il vient que (1) équivaut à (2). ◄

Si $\int \psi(\theta, y)\nu(dy)$ existe, on peut en qualité d'ensemble $A(t)$ pour les fonctions $\psi(t, y)$ continues $[\nu]$-presque partout prendre l'ensemble

$$A(t) = \{y: |\psi(t, y)| \leqslant 2|\psi(\theta, y)|\},$$

comme cela se fait, par exemple, dans la proposition suivante.

COROLLAIRE 1. *Supposons que $\pi(x)$ est une fonction mesurable bornée de $\mathscr{X}^n$ dans R et $f_\theta(x)$ une fonction continue par rapport à θ pour $[\mu^n]$-presque toutes les valeurs de $x \in \mathscr{X}^n$. Alors la fonction*

$$\mathsf{E}_\theta\pi(X) = \int \pi(x)f_\theta(x)\mu^n(dx)$$

est continue par rapport à θ.

Démonstration. Utilisons le théorème 1 pour $\mathscr{Y} = \mathscr{X}^n$, $y = x$, $\nu = \mu^n$, $\psi(t, x) = \pi(x)f_t(x)$, $A(t) = \{x : f_t(x) \leqslant 2f_\theta(x)\}$. Il est évident que les conditions 1) et 2) sont satisfaites. Vu que $\mathbf{E}_\theta\pi(X) \equiv 1$ est continue pour $\pi(x) \equiv 1$ on a (cf. (2))

$$\int_{x \notin A(t)} f_t(x)\mu^n(dx) \to 0$$

lorsque $t \to \theta$. D'après le théorème 1 on déduit de là que $\mathbf{E}_\theta\pi(X)$ est continue pour toute fonction bornée π. ◄

Si l'on ne s'intéresse qu'à une condition suffisante de convergence de (1) dans le cas où $\psi(t, y) \to \psi(\theta, y)$ p.p. pour $t \to \theta$, on peut prendre pour telle la convergence uniforme des intégrales de (1). Cette convergence peut être définie comme l'existence d'une mesure finie λ telle que l'inégalité $\lambda(A) < \delta = \delta(\varepsilon)$ entraîne $\sup_t \int_A |\psi(t, y)|\nu(dy) < \varepsilon$ pour $\varepsilon > 0$ donné.

S'il existe un majorant intégrable $\psi(y) = \sup_t \psi(t, y)$, cette mesure λ existe toujours : il suffit de poser $\lambda(A) = \int_A \psi(y)\nu(dy)$.

2. Conséquences des conditions (*R*). On prouvera ici le lemme 2.16.1 et la convergence uniforme de l'intégrale $I(\theta)$:

$$\sup_\theta \mathbf{E}_\theta(|l'(x_1, \theta)|^2;\ |l'(x_1, \theta)| > N) \to 0 \tag{3}$$

lorsque $N \to \infty$ (c'est justement cette uniformité qui est sous-entendue dans les §§ 2.24, 2.28 et 2.29). Dans ce numéro et dans le suivant, on n'envisage que le cas d'un paramètre scalaire, celui d'un paramètre vectoriel se traitant de façon analogue.

Théorème 2 (lemme 2.16.1). *Supposons que les conditions (R) sont remplies et que* $S = S(X)$ *est une statistique quelconque pour laquelle* $\mathbf{E}_\theta S^2 < c < \infty$, $\theta \in \Theta$. *Alors dans l'égalité*

$$a_S(\theta) = \mathbf{E}_\theta(S) = \int S(x)f_\theta(x)\mu^n(dx)$$

la dérivation sous le signe d'intégration est licite :

$$a_S'(\theta) = \int S(x)f_\theta'(x)\mu^n(dx) = \mathbf{E}_\theta SL'(X, \theta), \tag{4}$$

et de plus la fonction $a_S'(\theta)$ *est continue.*

Démonstration. Remarquons préalablement que pour $S(x) \equiv 1$ et $n = 1$ on déduit de (4) que

$$\int f_\theta'(x)\mu(dx) = 0. \tag{5}$$

Comme $L'(X, \theta) = \sum_{i=1}^{n} l'(x_i, \theta)$ est une somme de variables aléatoires indépendantes de moyenne nulle (cf. (5)), il vient

$$\mathbf{V}_\theta L'(X, \theta) = \mathbf{E}_\theta(L'(X, \theta))^2 = n\mathbf{E}_\theta(l'(x_1, \theta))^2 = nI(\theta). \tag{6}$$

Supposons maintenant que la fonction

$$I_n(\theta) = \mathbf{E}_\theta(L'(X, \theta))^2 = 4\int(\sqrt{f_\theta(x)}\,')^2\mu^n(dx)$$

est continue par rapport à θ (nous ne pouvons encore pas utiliser (6)). Appliquons le théorème 1 pour $\mathscr{Y} = \mathscr{X}^n$, $\nu = \mu^n$, $\psi(t, x) = (\sqrt{f_t(x)}')^2$, $\delta = t - \theta$,

$$A(t) = A_1(\delta) = \{x; \sup_{v:|\theta - v| < |v|} \sqrt{f_v(x)} < 2\sqrt{f_\theta(x)},$$
$$\sup_{v:|\theta - v| < |\delta|} |\sqrt{f_v(x)}'| \leq 2|\sqrt{f_\theta(x)}'|\}. \quad (7)$$

Les conditions 1) et 2) du théorème sont remplies pour $\psi(x) = 2\psi(\theta, x)$ puisque les fonctions $\sqrt{f_\theta}$ et $\sqrt{f_\theta}'$ sont continues. Donc, la convergence de $I_n(t)$ vers $I_n(\theta)$ lorsque $t \to \theta$ entraîne (cf. (2))

$$\epsilon(t) = \int_{x \notin A_1(\delta)} (\sqrt{f_t(x)}')^2 \mu^n(dx) \to 0, \quad t \to \theta. \quad (8)$$

Comme dans le corollaire 1, on en déduit la continuité de $\int S(x) f'_t(x) \mu^n(dx)$. Pour s'en assurer, il faut appliquer le théorème 1 « en sens inverse » pour les mêmes ensembles $A(t)$ et $\psi(t, x) = S(x) f'_t(x)$. Les conditions 1) et 2) du théorème 1 seront visiblement satisfaites ($\psi(x) = 2|S(x) f'_\theta(x)|$, $\int \psi(x)\mu^n(dx) \leq 4\mathrm{E}_\theta S^2 \int (\sqrt{f_\theta(x)}')^2 \mu^n(dx)$). La relation (2) a lieu en vertu de (8) et de l'inégalité, tout juste établie, dans laquelle l'intégration doit être effectuée sur l'ensemble des $x \notin A_1(\delta)$.

Prouvons maintenant (4). Remarquons que

$$\frac{1}{\delta}\left(\int S f'_{\theta+\delta}\mu^n - \int S f_\theta \mu^n\right) = \int \int_0^1 S f'_{\theta+u\delta} du \mu^n = \int \int_0^1 2S\sqrt{f_{\theta+u\delta}}(\sqrt{f_{\theta+u\delta}})' du \mu^n.$$

Appliquons de nouveau le théorème 1 pour $\mathscr{Y} = R \times \mathscr{X}^n$, $y = (u, x)$, $\nu = \lambda \times \mu^n$ (λ est la mesure de Lebesgue), $\psi(\delta, y) = S(x) f'_{\theta+u\delta}(x)$, $\delta \to 0$, $A(\delta) = A_1(\delta)$, où $A_1(\delta)$ est défini dans (7). La continuité de $\sqrt{f_\theta(x)}$ et $\sqrt{f_\theta(x)}'$ entraîne de nouveau les conditions 1) et 2) du théorème 1 :

$$\psi(\delta, y) I_{A(\delta)}(x) \to S(x) f'_\theta(x) = \psi(0, y) \quad \text{pour} \quad \delta \to 0,$$
$$\sup_\delta |\psi(\delta, y) I_{A(\delta)}(x)| \leq 4S(x)|f'_\theta(x)|,$$

où en vertu de l'inégalité de Cauchy-Bouniakovski

$$\int 4S|f'_\theta|\mu^n \leq 4\left[\int S^2 f_\theta \mu^n \cdot \int \frac{(f'_\theta)^2}{f_\theta}\mu^n\right]^{1/2} < \infty.$$

Pour établir (4) il faut donc vérifier la condition (2). Celle-ci résulte de l'inégalité de Cauchy-Bouniakovski et de la relation (8) :

$$\left|\int_{x \notin A_1(\delta)} \int_0^1 S\sqrt{f_{\theta+u\delta}}(\sqrt{f_{\theta+u\delta}})' du \mu^n\right| \leq$$
$$\leq \left[\int \int_0^1 S^2 f_{\theta+u\delta} du \mu^n\right]^{1/2} \left[\int_{x \notin A_1(\delta)} \int_0^1 (\sqrt{f_{\theta+u\delta}})'^2 du \mu^n\right]^{1/2} \leq$$
$$\leq c^{1/2}\left[\int_0^1 \epsilon(\theta + u\delta) du\right]^{1/2} \to 0$$

lorsque $\delta \to 0$.

Nous avons donc prouvé (4) sous la condition que $I_n(\theta)$ soit continue. Mais $I_n(\theta) = I(\theta)$ pour $n = 1$ et cette condition est remplie en vertu des conditions (R). Donc, la relation (4) est vraie pour $n = 1$ et, par suite, (5) l'est aussi. Mais (5) entraîne (6) qui exprime que $I(\theta)$ est continue. ◄

THÉORÈME 3. *Si* Θ *est compact et la fonction* $\sqrt{f_\theta(x)}$ *est continûment dérivable par rapport à* θ *pour* $[\mu]$*-presque toutes les valeurs de* x, *alors une condition nécessaire et suffisante pour que* $I(\theta)$ *soit continue est que* (3) *soit réalisée.*

Ce théorème exprime que dans la condition (R) la continuité de $I(\theta)$ peut être remplacée par la condition (3).

DÉMONSTRATION. Supposons que $I(\theta)$ est continue et que (3) n'est pas réalisée. Il existe alors un $\gamma > 0$ et des suites $t \to \theta \in M\mathfrak{G}$ $N_t \to \infty$, tels que

$$m(t) = \mathrm{E}_t[|l'(x_1, t)|^2; l'(x_1, t) > N_t] > \gamma \tag{9}$$

pour tous les t de la suite choisie.

Utilisons le théorème 1 pour $\mathscr{Y} = \mathscr{X}$, $\nu = \mu$, $\psi(t, x) = (\sqrt{f_t(x)}\,')^2 = \frac{1}{4}(l'(x, t))^2 f_t(x)$, $A(t) = \{x: |\sqrt{f_t(x)}\,'| \leqslant 2|\sqrt{f_\theta(x)}\,'|\}$. La fonction $\sqrt{f_\theta(x)}\,'$ étant continue, les conditions 1) et 2) du théorème 1 seront satisfaites et par suite la continuité de $I(t)$ entraînera

$$m_1(t) = \int_{x \notin A(t)} |\sqrt{f_t(x)}\,'|^2 \mu(dx) \to 0$$

lorsque $t \to \theta$. Mais $m(t) \leqslant m_1(t) + m_2(t)$, où

$$m_2(t) = \int_{B(t) \cap A(t)} (\sqrt{f_t}\,')^2 \mu, \quad B(t) = \{x: 2|\sqrt{f_t(x)}\,'| > N_t\sqrt{f_t(x)}\}.$$

De la forme de l'ensemble $A(t)$, il résulte que

$$m_2(t) \leqslant 4 \int_{B(t)} |\sqrt{f_\theta}\,'|^2 \mu.$$

En utilisant encore une fois la convergence $(\sqrt{f_t(x)})' \to (\sqrt{f_\theta(x)})'$, $\sqrt{f_t(x)} \to \sqrt{f_\theta(x)}$ lorsque $t \to \theta$, on trouve que $B(t)$ converge vers un ensemble μ-négligeable. Ceci exprime que $\mu(B(t)) \to 0$, $m_2(t) \to 0$, $m(t) \to 0$ lorsque $t \to \infty$. Cette contradiction avec (9) prouve (3).

Supposons maintenant que la relation (3) a lieu. D'après le théorème 1, pour établir la continuité de $I(t)$, il suffit de s'assurer que pour le même ensemble $A(t)$ que plus haut, on a $m_1(t) \to 0$ lorsque $t \to \infty$. Mais

$$m_1(t) \leqslant \int_{|l'|>N} |l'|^2 f_t \mu + N^2 \int_{x \notin A_1(t)} f_t \mu,$$

où la première intégrale peut être rendue arbitrairement petite, en vertu de (3), moyennant un choix convenable de N. Pour estimer la deuxième intégrale, on remarquera que $\mu(A(t)) \to 0$ et que pour $C(t) = \{x: f_t(x) \leqslant 2f_\theta(x)\}$ on a $\int_{x \notin C(t)} f_t \mu \to 0$ lorsque $t \to 0$ (cf. démonstration du corollaire 1). Donc

$$\int_{x \notin A(t)} f_t \mu \leqslant 2 \int_{x \notin A(t)} f_\theta \mu + \int_{x \notin C(t)} f_t \mu \to 0$$

lorsque $t \to \theta$. ◄

3. Conséquences des conditions (*RR*).

THÉORÈME 4. *Si les conditions* (*RR*) *sont réalisées, on a* $\int f\ddot{l}(x)\mu(dx) = 0$.

Combiné au théorème 2 ce théorème nous assure que les conditions (2.24.4) seront satisfaites.

DÉMONSTRATION. D'après le théorème 2, pour tous les $\theta \in \Theta$, on a

$$\int f\ddot{l}(x)\mu(dx) = 0$$

il nous suffit de prouver que pour $t \to \theta$

$$J(t) = \frac{1}{t-\theta}\left[\int f_t'\mu - \int f_\theta'\mu\right] \to \int f_\theta''\mu.$$

Remarquons que $\dfrac{1}{t-\theta}(f_t' - f_\theta') = \varphi_t f_t + \dfrac{f_\theta'}{f_\theta}\cdot\dfrac{f_t - f_\theta}{t-\theta}$, où $\varphi_t = \dfrac{1}{t-\theta}\left(\dfrac{f_t'}{f_t} - \dfrac{f_\theta'}{f_\theta}\right)$.

En se servant de cette égalité, on peut représenter $J(t)$ par la somme des quatre termes $J(t) = J_1 + J_2 + J_3 + J_4$, où

$$J_1 = \int \varphi_t f_\theta \mu, \qquad J_2 = \int_{l \leqslant N} \varphi_t (f_t - f_\theta)\mu,$$

$$J_3 = \int_{l \geqslant N} \varphi_t (f_t - f_\theta)\mu, \quad J_4 = \int \frac{f_\theta'}{f_\theta}\cdot\frac{f_t - f_\theta}{t-\theta}\mu,$$

$l = l(x)$ étant le majorant de $l''(x, t)$ dans les conditions (*RR*). D'après le théorème 2, il vient pour $n = 1$, $S(x) = l'(x, \theta)$

$$J_4 = \frac{1}{t-\theta}(\mathrm{E}_t l'(x_1, \theta) - \mathrm{E}_\theta l'(x_1, \theta)) \to \mathrm{E}_\theta (l'(x_1, \theta))^2 = I(\theta). \tag{10}$$

Par ailleurs

$$|\varphi_t| < l \tag{11}$$

et par suite en vertu du théorème de Lebesgue

$$\lim_{t\to\theta} J_1 = \int \lim_{t\to\theta} \varphi_t f_\theta \mu = \int l'' f_\theta \mu = \int f_\theta'' \mu - I(\theta). \tag{12}$$

En se servant encore de (11), on trouve en vertu des conditions (*RR*)

$$|J_3| \leqslant \int_{l \geqslant N} l f_t \mu + \int_{l \geqslant N} l f_\theta \mu \to 0$$

lorsque $N \to \infty$. L'inégalité de Cauchy-Bouniakovski nous donne enfin

$$|J_2| \leqslant N\int |f_t - f_\theta|\mu \leqslant N\int\int_\theta^t |f_u'|du\mu \leqslant N\int_\theta^t \sqrt{I(u)}\,du \to 0 \tag{13}$$

lorsque $t \to \theta$. En combinant les relations (10) à (13), on trouve que $0 = J(t) \to \int f_\theta''\mu$. ◄

ANNEXE VII

INÉGALITÉS POUR LA DISTRIBUTION DU RAPPORT DE VRAISEMBLANCE DANS LE CAS MULTIDIMENSIONNEL

Dans ce numéro on prouvera le théorème suivant (théorème 28.2 ; les notations sont celles des §§ 2.21, 2.23 et 2.28).

THÉORÈME 1. *Supposons remplies les conditions suivantes :*

$$\inf_{u} \frac{r(u)}{u^2} \geqslant g(\theta) > 0, \tag{1}$$

$$\mathbf{E}_\theta l'(x_1, \theta) = 0, \tag{2}$$

$$\gamma = \sup_\theta \mathbf{E}_\theta |l'(x_1, \theta)|^s < \infty \tag{3}$$

pour un certain $s > k$. Alors pour tous z, $n \geqslant 1$

$$\mathbf{P}_\theta(\sup_{|v| \geqslant r} Z(v/\sqrt{n}) \geqslant e^z) \leqslant c\gamma(e^{-z/2} + e^{-z})e^{-\beta g(\theta) r^2},$$

où $\beta > 0$ dépend seulement de k et de s ; $c < \infty$ dépend de k, s et de $g(\theta)$ et peut être choisi indépendant de $g(\theta)$ si $g(\theta) > g > 0$ pour tous les θ.

Nous aurons besoin de quelques propositions auxiliaires. Par c_s et $c_{k,s}$ nous désignerons des constantes qui dépendront seulement de leurs indices.

LEMME 1. *Soient ξ_k, $k = 1, 2, \ldots$, des variables aléatoires indépendantes et équidistribuées, $\mathbf{E}\xi_1 = 0$, $\mathbf{E}|\xi_1|^s < \gamma < \infty$, $s \geqslant 2$. Alors*

$$\mathbf{E}\left|\sum_{k=1}^{n} \xi_k\right|^s \leqslant c_s \gamma n^{s/2}.$$

DÉMONSTRATION. Pour alléger les raisonnements on se limitera au cas où $s = 2m$ est un entier pair *). Dans ce cas, il suffit d'envisager des variables aléatoires scalaires ξ_k. On a

$$\mathbf{E}\left|\sum_{k=1}^{n} \xi_k\right|^s = \sum_{k_1, \ldots, k_n} \mathbf{E}\xi_1^{k_1} \ldots \mathbf{E}\xi_n^{k_n}, \tag{4}$$

où la sommation est étendue à tous les entiers $k_1, \ldots, k_n$ tels que $\sum_j k_j = s$, $k_j \neq 1$ (les $k_j =$ $= 1$ sont exclus, car $\mathbf{E}\xi_k = 0$). L'inégalité de Hölder nous donne

$$|\mathbf{E}\xi_j^{k_j}| \leqslant (\mathbf{E}|\xi_j|^s)^{k_j/s} = \gamma^{k_j/s}$$

---

*) La démonstration dans le cas général est accessible par exemple dans [42].

donc
$$\prod_{j=1}^{n} |\mathbf{E}\xi_j^{k_j}| \leqslant \prod_{j=1}^{n} \gamma^{k_j/s} = \gamma.$$

Reste à estimer $\sum_{k_1,\ldots,k_n} 1$. Désignons par $(l_1, \ldots, l_p)$ les éléments non nuls ($l_i \geqslant 2$) de l'ensemble $(k_1, \ldots, k_n)$ $\left(\sum_{j=1}^{p} l_j = s\right)$. La somme estimée sera égale à $\sum_{(l_1,\ldots,l_p)} A_p$, A_p est le nombre d'arrangements des éléments $l_1, \ldots, l_p$ pris n à n. Il est évident que $A_p \leqslant n(n-1)\ldots$ $\ldots(n-p+1)$. La plus grande valeur de p est égale à $m = s/2$ (elle correspond à l'ensemble $(2, 2, \ldots, 2)$), de sorte que $A_p \leqslant A_m \leqslant n^m$. Or le nombre des arrangements $(l_1, \ldots, l_p)$ ne dépend que de s. Donc la somme estimée est au plus égale à $c_s n^m$. ◄

Posons $p(u) = Z^{1/s}(u)$.

LEMME 2. *Si les conditions* (2), (3) *sont réalisées, on a*
$$\mathbf{E}_\theta |p'(u)|^s \leqslant c_s \gamma n^{s/2}.$$

DÉMONSTRATION.
$$\mathbf{E}_\theta |p'(u)|^s = \mathbf{E}_\theta \left| \frac{1}{s} L'(X, \theta + u) Z^{1/s}(u) \right|^s =$$
$$= s^{-s}\mathbf{E}_\theta |L'(X, \theta + u)|^s Z(u) = s^{-s}\mathbf{E}_{\theta+u} |L'(X, \theta + u)|^s.$$

Reste à appliquer le lemme 1 aux variables aléatoires $\xi_k = l'(x_k, \theta + u)$. ◄

LEMME 3. *Si les conditions du théorème* 1 *sont remplies, on a*
$$\mathbf{E}_\theta |p(u+v) - p(u)|^s \leqslant |v|^s c_s \gamma n^{s/2},$$
où c_s est le même que dans le lemme 2.

DÉMONSTRATION. L'inégalité de Hölder et le lemme 2 nous donnent
$$\mathbf{E}_\theta |p(u+v) - p(u)|^s = \mathbf{E}_\theta \left| \int_0^{|v|} (p'(u + tv/|v|), v/|v|)dt \right|^s =$$
$$= |v|^s \mathbf{E}_\theta \left| \int_0^1 (p'(u + hv), v/|v|)dh \right|^s \leqslant |v|^s \int_0^1 \mathbf{E}_\theta |p'(u + hv)|^s dh \leqslant |v|^s c_s \gamma n^{s/2}. \quad ◄$$

Désignons par $K_{u,\Delta}$ un cube de R^k d'arête Δ et de sommet $u = (u_1, \ldots, u_k)$:
$$K_{u,\Delta} = \{v \in R^k : u_i \leqslant v_i \leqslant u_i + \Delta,\ i = 1, \ldots, k\}.$$

LEMME 4. *Si les conditions du théorème* 1 *sont remplies, on a*
$$\mathbf{P}_\theta\left(\sup_{v \in K_{u,\Delta}} Z(v/\sqrt{n}) \geqslant e^z\right) \leqslant c\gamma\Delta^k (e^{-z/2} + e^{-z}) e^{-|u|^2 \beta g(\theta)},$$
où les constantes $c < \infty$ et $\beta > 0$ ne dépendent que de k et s,
$$\Delta = e^{-|u|^2 g(\theta)/(4k)}.$$

Cette estimation sera valable pour tout cube d'arête Δ contenant le point u.

DÉMONSTRATION. Représentons le point $v \in K_{u,\Delta}$ sous la forme $v = u + t\Delta$, où $t \in K_{0,1}$. Utilisons le développement binaire des coordonnées t_i du vecteur t :

$$t_i = \sum_{r=1}^{\infty} \frac{\delta_{ir}}{2^r},$$

où les δ_{ir} sont égaux à 0 ou à 1. Posons

$$t_i^m = \sum_{r=1}^{m} \frac{\delta_{ir}}{2^r}, \quad t^m = (t_1^m, \ldots, t_k^m), \quad t^0 = 0, \tag{5}$$

de sorte que t^m est une approximation binaire de t : $|t - t^m| < 2^{-m}\sqrt{k}$.

Désignons $\varphi(t) = p\left(\frac{v}{\sqrt{n}}\right) = p\left(\frac{u + t\Delta}{\sqrt{n}}\right)$. Alors

$$\varphi(t) = \varphi(0) + \sum_{m=1}^{\infty} (\varphi(t^m) - \varphi(t^{m-1})).$$

On dira que des points $t^m_{(1)}$ et $t^m_{(2)}$ sont voisins si leurs représentations (5) ne diffèrent que par un seul nombre $\delta_{1m}, \ldots, \delta_{km}$. Il est clair que si pour deux points voisins quelconques on a

$$|\varphi(t^m_{(1)}) - \varphi(t^m_{(2)})| < c_m/\sqrt{k}, \tag{6}$$

alors $|\varphi(t^m) - \varphi(t^{m-1})| < c_m$. Donc, si (6) a lieu pour tous les points voisins quels que soient m et $c_m = a(1 - q)q^m$, $q < 1$, et si de plus

$$\varphi(0) < a(1 - q), \tag{7}$$

alors

$$\varphi(t) < \sum_{m=0}^{\infty} a(1 - q)q^m = a,$$
$$\sup_{t \in K_{0,1}} \varphi(t) < a. \tag{8}$$

Considérons maintenant la proposition du lemme. Il nous faut estimer

$$\mathbf{P}_\theta\Big(\sup_{v \in K_{u,\Delta}} Z(v/\sqrt{n}) \geqslant e^z\Big) = \mathbf{P}_\theta\Big(\sup_{t \in K_{0,1}} \varphi(t) \geqslant a\Big) \tag{9}$$

pour $a = e^{z/s}$. L'inégalité se trouvant sous le signe de la probabilité met en défaut la relation (8) et partant l'une des inégalités (6), (7). La probabilité (9) est donc estimée par la somme des probabilités

$$\mathbf{P}_\theta(\varphi(0) \geqslant a(1 - q)) + \sum_{m=1}^{\infty} \sum \mathbf{P}_\theta\left(|\varphi(t^m_{(1)}) - \varphi(t^m_{(2)})| \geqslant \frac{a(1 - q)q^m}{\sqrt{k}}\right), \tag{10}$$

où la dernière somme est formée par $k(2^m)^k$ termes associés à tous les couples possibles de points voisins $t^m_{(1)}$ et $t^m_{(2)}$. Comme $|t^m_{(1)} - t^m_{(2)}| = 2^{-m}$, pour chacun de ces couples, le lemme 3 et l'inégalité de Tchébychev nous donnent

$$\mathbf{P}_\theta\left(|\varphi(t^m_{(1)}) - \varphi(t^m_{(2)})| \geqslant \frac{a(1 - q)q^m}{\sqrt{k}}\right) \leqslant \left(\frac{\Delta}{\sqrt{n}2^m}\right)^s c_s \gamma n^{s/2} \left(\frac{a(1 - q)q^m}{\sqrt{k}}\right)^{-s}.$$

Donc, la double somme de (10) sera majorée par

$$\Delta^s c_s \gamma [a(1 - q)]^{-s} k^{1-s/2} \sum_{m=1}^{\infty} 2^{-m(s-k)} q^{-ms}. \tag{11}$$

Choisissons $q = 2^{\frac{k-s}{2s}} < 1$. La série de (11) sera alors convergente et l'expression (11) sera de la forme $c_{k,s}\gamma\Delta^s a^{-s}$. En vertu du théorème 2.28.1 et de l'inégalité de Tchébychev, on a pour le premier terme de (10)

$$\mathbf{P}_\theta(\varphi(0) \geqslant a(1 - q)) = \mathbf{P}_\theta(Z^{1/2}(u/\sqrt{n}) \geqslant (a(1 - q))^{s/2}) \leqslant (a(1 - q))^{-s/2} e^{-|u|^2 g(\theta)/2}.$$

En posant $\Delta = e^{-|u|^2 g(\theta)/(4k)}$, en tenant compte de ce que $a^{-s} = e^{-z}$ et en admettant sans perdre en généralité que $s \leqslant 2k$, on trouve que (9) est majorée par

$$c_{k,s}\gamma\Delta^k e^{-|u|^2 g(\theta)(s-k)/(4k)}(e^{-z} + e^{-z/2}).$$

Pour $\beta = \dfrac{s-k}{4k}$, on obtient la première proposition du lemme.

La véracité de la deuxième proposition du lemme est évidente, puisque dans la démonstration on aurait pu remplacer $\varphi(0)$ par la valeur de la fonction $\varphi(t_0)$ en un point fixe quelconque $t_0 \in K_{0,1}$ (la première somme de (10) correspond à la valeur prise en un point, la deuxième, à la variation totale éventuelle de $\varphi(t)$ dans $K_{0,1}$). ◄

DÉMONSTRATION du théorème 1. Recouvrons l'espace R^k tout entier par un système de cubes $K_{u,\Delta}$ dont les coordonnées des points u sont des multiples de Δ. Les cubes coupant $S_r = \{v \in R^k : r \leqslant |v| \leqslant r + 1\}$ sont en nombre inférieur à $c_k r^{k-1}\Delta^{-k}$. Donc

$$\mathbf{P}_\theta\Big(\sup_{v \in S_r} Z(v/\sqrt{n}) \geqslant e^z\Big) \leqslant c_k r^{k-1} c_{k,s}\gamma(e^{-z/2} + e^{-z})e^{-r^2\beta g(\theta)},$$

$$\mathbf{P}_\theta\Big(\sup_{|v| \geqslant r} Z(v/\sqrt{n}) > e^z\Big) \leqslant c_k c_{k,s}\gamma(e^{-z/2} + e^{-z}) \sum_{j=0}^{\infty} (r + j)^{k-1} e^{-(r+j)^2 \beta g(\theta)}.$$

Comme $\sup_{r,j}(r + j)^{k-1} e^{-\beta g(\theta)(r+j)^2/2}$ et $\sum_j e^{-\beta g(\theta)[(r+j)^2 - r^2]/2} < \sum_j e^{-\beta g(\theta) j^2/2}$ sont majorées par une constante $c(\beta g(\theta))$ ne dépendant que de $\beta g(\theta)$, l'expression obtenue est au plus égale à $c_{k,s}\gamma c(\beta g(\theta)) e^{-r^2 \beta g(\theta)/2}(e^{-z/2} + e^{-z})$, où $c(\beta g(\theta)) < c(\beta g) < \infty$ si $g(\theta) \geqslant g > 0$ pour tous les θ. ◄

ANNEXE VIII

DÉMONSTRATION DE DEUX THÉORÈMES FONDAMENTAUX DE LA THÉORIE DES JEUX STATISTIQUES

On admettra que sont remplies les conditions suivantes.

CONDITION (A). *L'ensemble D des décisions et l'ensemble Θ des paramètres (des stratégies pures de la nature) sont des espaces métriques compacts munis des métriques respectives ϱ_D et ϱ_Θ.*

CONDITION (B). *La fonction de perte $w(\delta, \theta) : D \times \Theta \to R$ est continue par rapport à δ et θ pour les métriques ϱ_D et ϱ_Θ respectivement.*

Nous nous passerons de la condition $w(\delta, \theta) \geqslant 0$.

Nous disposons d'un échantillon $X \in \mathbf{P}_\theta$ dont la taille peut être supposée égale à 1 sans perte en généralité.

CONDITION (C). *Les distributions $\mathbf{P}_\theta$ sont continues en variation par rapport à θ, c'est-à-dire que*

$$\sup_{B \in \mathfrak{B}_{\mathscr{X}}} |\mathbf{P}_{\theta_m}(B) - \mathbf{P}_\theta(B)| \to 0,$$

si $\varrho_\Theta(\theta_m, \theta) \to 0$ lorsque $m \to \infty$.

Si la condition (A_μ) est remplie, c'est-a-dire si $\mathbf{P}_\theta$ admet une densité $f_\theta(x)$ par rapport à une mesure σ-finie μ sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}})$:

$$f_\theta(x) = \frac{d\mathbf{P}_\theta}{d\mu}(x),$$

alors la condition (C) équivaudra à la continuité de $f_\theta(x)$ dans $L_1(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}, \mu)$:

$$\int |f_{\theta_m}(x) - f_\theta(x)| \mu(dx) \to 0,$$

si $\varrho_\Theta(\theta_m, \theta) \to 0$ lorsque $m \to \infty$.

Les conditions (A), (B) et (C) sont de toute évidence valables pour des ensembles D et Θ finis.

Si D est composé d'un nombre fini de points $\delta_1, \ldots, \delta_r$, la condition (A) relative à D sera remplie (le choix de ϱ_D ne joue aucun rôle), quant à la condition (B), elle exprimera la continuité par rapport à ϱ_Θ des fonctions $w(\delta_1, \theta), \ldots, w(\delta_r, \theta)$.

Si les ensembles D et Θ sont tous deux finis, les conditions (A), (B) et (C) sont automatiquement satisfaites.

Désignons par σ_D et σ_Θ les tribus respectives des boréliens de D et de Θ. Suivant le § 2.3, on désignera par $(\tilde{\mathscr{D}}, \tilde{\Theta}, \tilde{W})$ le jeu statistique moyennisé, dont les éléments de $\tilde{\Theta}$ sont des distributions $\mathbf{Q}$ sur (Θ, σ_Θ) et les éléments de $\tilde{\mathscr{D}}$ des distributions $\pi(x) = \pi(x, \cdot)$ sur (D, σ_D) (pour tout $x \in \mathscr{X}$), où $\pi(x, A)$ est une fonction mesurable par rapport à x quel que soit $A \in \sigma_D$.

La fonction de risque $\tilde{W}(\pi, Q)$ est définie par

$$\tilde{W}(\pi, Q) = \int_{\Theta}\int_{\mathscr{T}}\int_{D} w(u, t)\pi(x, du)f_t(x)\mu(dx)Q(dt).$$

Si l'on remplace l'argument Q par θ, la fonction $\tilde{W}(\pi, \theta)$ deviendra $\tilde{W}(\pi, \mathbf{I}_\theta)$, où $\mathbf{I}_\theta$ est une distribution concentrée au point θ. Cette convention sera également valable si l'on remplace $\pi \in \tilde{\mathscr{D}}$ par $\delta \in \mathscr{D}$. Il nous sera également commode de substituer W à $\tilde{W}$. Aucune confusion ne sera à craindre.

LEMME 1. *Si les conditions* (A), (B) *et* (C) *sont satisfaites, la fonction* $W(\pi, \theta)$ *est continue par rapport à* θ *pour toute stratégie* $\pi(x)$.

DÉMONSTRATION. Lorsque $\theta_n \to \theta$ on a

$$\begin{aligned}|W(\pi, \theta_n) - W(\pi, \theta)| &\leqslant |\mathbf{E}_\theta \mathbf{E}[w(\pi(X), \theta) - w(\pi(X), \theta_n)|\, X]| + \\ &+ |\mathbf{E}_\theta \mathbf{E}[w(\pi(X), \theta_n)|X] - \mathbf{E}_{\theta_n}\mathbf{E}[w(\pi(X), \theta_n)|X]| \leqslant \\ &\leqslant \int|w(\pi(x), \theta) - w(\pi(x), \theta_n)|\mathbf{P}_\theta(dx) + \sup_{\delta,\theta}|w(\delta, \theta)|\int|\mathbf{P}_{\theta_n}(dx) - \mathbf{P}_\theta(dx)|. \qquad (1)\end{aligned}$$

La première intégrale converge ici vers 0 en vertu de la continuité de la fonction w par rapport à θ. La convergence vers 0 de la deuxième intégrale résulte de la condition (C). En effet, supposons que $f_{\theta_n}(x)$ est la densité de $\mathbf{P}_{\theta_n}$ par rapport à la mesure

$$\mu = \mathbf{P}_\theta + \sum_{j=1}^{\infty} 2^{-j}\mathbf{P}_{\theta_j}.$$

et que $B_n = \{x: f_{\theta_n}(x) \geqslant f_\theta(x)\}$. La deuxième intégrale de (1) est alors égale à

$$\int|f_{\theta_n}(x) - f_\theta(x)|\mu(dx) = 2\int_{B_n}(f_{\theta_n}(x) - f_\theta(x))\mu(dx) = 2(\mathbf{P}_{\theta_n}(B_n) - \mathbf{P}_\theta(B_n)) \to 0. \quad ◄$$

THÉORÈME 1 (premier théorème fondamental). *Si les conditions* (A) (B) *et* (C) *sont remplies, le jeu* $(\tilde{\mathscr{D}}, \tilde{\Theta}, W)$ *admet une valeur et des stratégies minimax pour les deux joueurs. En d'autres termes, il existe une distribution la plus défavorable* $\overline{Q}$ *et une décision minimax* $\bar{\pi}(x)$:

$$W_* = \sup_Q \inf_\pi W(\pi, Q) = W(\bar{\pi}, \overline{Q}) = \inf_\pi \sup_Q W(\pi, Q) = W^*. \qquad (2)$$

En vertu du lemme 2.1 la proposition (2) équivaut à :

$$W(\bar{\pi}, \uparrow) \equiv \sup_Q W(\bar{\pi}, Q) = W(\bar{\pi}, \overline{Q}) = \inf_\pi W(\pi, \bar{Q}) \equiv W(\downarrow, \bar{Q}). \qquad (3)$$

THÉORÈME 2 (deuxième théorème fondamental). *Si les conditions* (A), (B) *et* (C) *sont remplies, les décisions bayésiennes* $\pi_Q(x)$ *forment une classe complète. En d'autres termes, pour tout* $\pi_0 \in \tilde{\mathscr{D}}$ *il existe un* $Q \in \tilde{\Theta}$ *et un* $\pi_Q \in \tilde{\mathscr{D}}$ *tels que*

1) $W(\pi_Q, Q) = W(\downarrow, Q)$,

2) $W(\pi_Q, \theta) \leqslant W(\pi_0, \theta), \forall\theta$.

DÉMONSTRATION du théorème 2. Le deuxième théorème fondamental découle du premier. Considérons une stratégie quelconque $\pi_0 \in \tilde{\mathscr{D}}$ et le jeu $(\tilde{\mathscr{D}}, \tilde{\Theta}, W_0)$, où W_0 est construite à l'aide de la fonction $w_0(\delta, \theta) = w(\delta, \theta) - W(\pi_0, \theta)$, de sorte que

$$W_0(\pi, \theta) = W(\pi, \theta) - W(\pi_0, \theta). \qquad (4)$$

La fonction $v(\theta) = W(\pi_0, \theta)$ est continue par rapport à θ en vertu du lemme 1, donc, la fonction de perte $w_0(\delta, \theta) = w(\delta, \theta) - v(\theta)$ et $w(\delta, \theta)$ satisfont la condition (B). Ceci exprime que le jeu $(\mathscr{D}, \bar{\Theta}, W_0)$ est justiciable du théorème 1. Comme $W_0(\pi_0, \uparrow) = 0$ (cf. (4)), la valeur supérieure de ce jeu vérifie l'inégalité $W\bar{\delta} \leqslant 0$. De (2) et (3), il s'ensuit alors qu'il existe des $\bar{\pi}$ et $\overline{Q}$ tels que

$$\sup_{P} W_0(\bar{\pi}, P) = \sup_{\theta} W_0(\bar{\pi}, \theta) \leqslant 0, \ \bar{\pi} = \pi_{\overline{Q}}.$$

Ces deux relations sont équivalentes aux propositions 2) et 1) du théorème 2 si l'on pose $\overline{Q} = Q$ et $\bar{\pi} = \pi_Q$. ◄

Le théorème 1 résulte des deux lemmes suivants.

LEMME 2. *Si les conditions* (A), (B) *et* (C) *sont réalisées, il existe une distribution* $\overline{Q}$ *telle que* $W(\downarrow, \overline{Q}) \geqslant \inf_{\pi} W(\pi, \uparrow) = W^*$.

LEMME 3. *Si les conditions* (A), (B) *et* (C) *sont satisfaites, il existe une stratégie* $\bar{\pi}$ *telle que* $W(\bar{\pi}, \uparrow) \leqslant W^*$.

Les inégalités des lemmes 2 et 3 etraînent la relation

$$W^* \geqslant W(\bar{\pi}, \uparrow) \geqslant W(\bar{\pi}, \overline{Q}) \geqslant W(\downarrow, \overline{Q}) \geqslant W^*,$$

qui est équivalente à (3) et par suite à (2). Ce qui prouve le théorème 1. ◄

Les lemmes 2 et 3 divisent la démonstration du théorème 1 en deux parties. La première (lemme 2) est très peu liée au fait que le jeu est statistique. Cette partie de la démonstration se déroule à peu de choses près comme pour les jeux ordinaires (comparer avec [25]).

DÉMONSTRATION du lemme 2. Soit V l'ensemble des fonctions $\Theta \to R$ de la forme $v(\theta) = W(\pi, \theta)$, $\pi \in \mathscr{D}$. Le lemme 1 nous dit que toutes les fonctions de V sont continues, de sorte que $V \subset C(\Theta)$, où $C(\Theta)$ est l'ensemble de toutes les fonctions continues sur Θ. Soient par ailleurs $v_1(\theta) = W(\pi_1, \theta)$, $v_2(\theta) = W(\pi_2, \theta)$. Comme

$$v(\theta) = pv_1(\theta) + (1 - p)v_2(\theta) = W(p\pi_1 + (1 - p)\pi_2, \theta),$$

$$\pi = p\pi_1 + (1 - p)\pi_2 \in \mathscr{D},$$

pour $p \in [0, 1]$, il s'ensuit que $v \in V$ et par suite l'ensemble V est convexe.

Remarquons maintenant que $W^* = \inf_{\pi} W(\pi, \uparrow) = \inf_{v \in V} \sup_{\theta} v(\theta)$. Pour des raisons de commodité on envisagera la fonction $\dfrac{w(\delta, \theta) - v_0 + 1}{W^* - v_0 + 1}$, $v_0 = \inf_{v \in V} \inf_{\theta} v(\theta)$, au lieu de la fonction initiale $w(\delta, \theta)$. En désignant la nouvelle fonction encore par $w(\delta, \theta)$ (le problème reste le même) on obtient

$$W^* = 1, \quad v_0 > 0. \tag{5}$$

Soit maintenant U l'ensemble des fonctions continues $v(\theta) : \Theta \to R$ telles que $\sup_{\theta} v(\theta) < 1$. Il est évident que U est un ensemble convexe ouvert de $C(\Theta)$. D'autre part, la relation 5) entraîne la vacuité de l'intersection $V \cap U$. Donc, en vertu du théorème de Hahn-Banach cf. par exemple [25]) il existe une fonctionnelle linéaire $L(v) : C(\Theta) \to R$ telle que

$$L(v) < 1 \quad \text{si } v \in U, \quad L(v) \geqslant 1 \quad \text{si } v \in V. \tag{6}$$

Cette fonctionnelle possède nécessairement la propriété suivante : $L(v) \geqslant 0$ si $v(\downarrow) = \inf_{\theta} v(\theta) \geqslant 0$. En effet, si l'on admet qu'il existe un élément $v_0 \in C(\Theta)$, $v_0(\downarrow) \geqslant 0$, pour lequel

$L(v_0) < 0$, on trouve que $v_s = -sv_0 \in U$ pour tout $s > 0$, $L(v_s) = -sL(v_0) > 1$ pour s assez grand. Ce qui contredit (6).

Mais le théorème de Riesz ([36]) nous dit que la fonctionnelle positive L se représente par l'intégrale

$$L(v) = \int_\Theta v(\theta)\lambda(d\theta),$$

où λ est une mesure finie. Vu que $1 \geqslant \sup_{v\in U} L(v) = \lambda(\Theta)$, en admettant que $\overline{Q}(A) = \lambda(A)/\lambda(\Theta)$, on trouve pour $v \in V$ que :

$$L(v) = \int W(\pi, \theta)\lambda(d\theta) = \lambda(\Theta)W(\pi, \overline{Q}),$$

$$W(\downarrow, \overline{Q}) = \frac{1}{\lambda(\Theta)} \inf_{v\in V} L(v) \geqslant 1 = W^*. \blacktriangleleft$$

Démonstration du lemme 3. La fonction $W(\pi, \theta)$ étant continue par rapport à θ pour chaque $\pi \in \mathcal{D}$ (cf. lemme 1), il nous suffit de construire une stratégie $\bar{\pi}$ telle que pour tous les $k = 1, 2, \ldots$

$$W(\bar{\pi}, \theta_k) \leqslant W^*, \tag{7}$$

où θ_k sont des points d'un ensemble $T = \{\theta_1, \theta_2, \ldots\}$ dénombrable partout dense dans D. Par définition de la valeur supérieure W^* d'un jeu il existe une suite de stratégies $\pi_n = \pi_n(x, \cdot)$ telle que

$$W(\pi_n, \theta_k) < W^* + 1/n \tag{8}$$

pour tous les k.

Construisons maintenant à l'aide des distributions π_n une suite d'éléments aléatoires ζ_n spécialement choisis et extrayons d'elle une sous-suite convergente. A cet effet, désignons par $f_{\theta_k}(x)$ la densité de la distribution $\mathbf{P}_{\theta_k}$ par rapport à la mesure de probabilité $\mu = \sum_{j=1}^{\infty} 2^{-k}\mathbf{P}_{\theta_k}$, si bien que

$$W(\pi_n, \theta_k) = \iint w(u, \theta_k)\pi_n(x, du)f_{\theta_k}(x)\mu(dx). \tag{9}$$

Considérons l'espace $D\times R^T$, où R^T est l'espace des valeurs des éléments $f(x) = \{f_{\theta_1}(x), f_{\theta_2}(x), \ldots\}$ muni de la tribu $\mathfrak{B}^T$ engendrée par les ensembles cylindriques. Associons à chaque stratégie π un espace probabilisé $(D \times \mathscr{X}; \sigma_D \times \mathfrak{B}_{\mathscr{X}}; \mathbf{P})$, où la distribution $\mathbf{P}$ est définie par

$$\mathbf{P}(\delta \in A, X \in B) = \int_B \mu(dx)\pi(x, A), \quad (A \in \sigma_D, A \in \mathfrak{B}_{\mathscr{X}} \tag{10}$$

Définissons sur cet espace les éléments aléatoires $\zeta = \zeta(\delta; X) = (\delta; f_{\theta_1}(X), f_{\theta_2}(X), \ldots) = (\delta; f(X))$ et désignons par ζ_n les éléments associés à π_n, si bien que ζ_n sont des variables aléatoires sur l'espace probabilisé $(D \times R^T, \sigma_D \times \mathfrak{B}^T, \Pi_n)$ et la distribution Π_n est engendrée par π_n, la formule (10) et l'application $\zeta(\delta, x) : D \times \mathscr{X} \to D \times R^T$.

Désignons par $\Pi_n^{(k)}$ les restrictions de la distribution Π_n à $D \times R^k$ (ceci est la distribution conjointe de $(\delta; f_{\theta_1}(X), \ldots, f_{\theta_k}(X))$ et par λ la distribution de $f(X)$ sur $(\mathscr{X}, \mathfrak{B}_{\mathscr{X}}; \mu)$. Nous aurons besoin du

Lemme 4. *Il existe une distribution $\overline{\Pi}$ sur l'espace mesurable $(D \times R^T, \sigma_D \times \mathfrak{B}^T)$ et une sous-suite $\{\pi_{n^*}\} \subset \{\pi_n\}$ telles que*

$$\Pi_{n^*}^{(k)} \Rightarrow \overline{\Pi}^{(k)} \tag{11}$$

pour tout k ($\overline{\Pi}^{(k)}$ sont les restrictions de $\overline{\Pi}$),

$$\overline{\Pi}(D \times C) = \lambda(C), \quad C \in \mathfrak{B}^T. \tag{12}$$

Ce lemme sera prouvé plus bas.

Désignons par $\bar{\zeta} = (\bar{\delta}; \bar{f})$ l'élément aléatoire de distribution $\overline{\Pi}$. La relation (12) exprime que la distribution de $\bar{f}$ est confondue avec λ (lorsque n varie la deuxième « coordonnée » de ζ_n ne modifie pas la distribution). L'espace D étant métrique et compact, donc séparable, il existe (cf. [34]) une distribution conditionnelle (régulière) de $\bar{\delta}$ par rapport à $\bar{f}(X)$ que nous désignerons par $\Pi(\cdot|\bar{f}(x))$.

Considérons la stratégie $\bar{\pi}(x, A) = \Pi(\bar{\delta} \in A|\bar{f}(X))$ et montrons qu'elle est justiciable de (7).

Remarquons préalablement que

$$\mathsf{E}w(\bar{\delta}, \theta_k)\bar{f}_{\theta_k} = \mathsf{E}\bar{f}_{\theta_k}\mathsf{E}(w(\bar{\delta}, \theta_k)|X) = \int f_{\theta_k}(x)\int w(u, \theta_k)\bar{\pi}(u, dx)\mu(dx) = W(\bar{\pi}, \theta_k). \tag{13}$$

Le lemme 4 nous apprend que la distribution de $(\delta_{n^*}, f_{\theta_k}(X))$ converge faiblement vers celle de $(\bar{\delta}, \bar{f}_{\theta_k}(X))$. Puisque la fonction w est continue, la distribution conjointe de $(w(\delta_{n^*}, \theta_k), f_{\theta_k}(X))$ converge faiblement vers celle de $(w(\bar{\delta}, \theta_k), \bar{f}_{\theta_k}(X))$. Mais la fonction $g(u, v) = w(u, \theta_k)v$ est continue par rapport à u et v et est majorée par une fonction $g(v) = cv$, $c = \max_u w(u, \theta_k)$ telle que $\mathsf{E}g(f_{\theta_k}(X)) = c\int f_{\theta_k}(x)\mu(dx) = c$. Donc, d'après le théorème de continuité des moments (cf. théorème 1.5.4)

$$\lim_{n^* \to \infty} \mathsf{E}g\left(\delta_{n^*}, f_{\theta_k}(X)\right) = \mathsf{E}g(\bar{\delta}, \bar{f}_{\theta_k}(X)),$$

ou ce qui est équivalent $\lim_{n^* \to \infty} \mathsf{E}w(\delta_{n^*}, \theta_k)f_{\theta_k}(X) = \mathsf{E}w(\bar{\delta}, \theta_k)\bar{f}_{\theta_k}(X)$.

En vertu de (9) et (13) ceci nous conduit à la convergence

$$\lim_{n^* \to \infty} W(\pi_{n^*}, \theta_k) = W(\bar{\pi}, \theta_k).$$

Ce qui prouve le lemme 3, puisque le premier membre de cette égalité (cf. (8)) est au plus égal à W^*.

Démonstration du lemme 4. Figeons un $k \geqslant 1$ quelconque et traitons $D \times R^k$ comme un espace séparable métrique complet pour la métrique engendrée par la métrique euclidienne de R^k et la métrique ϱ_D. Pour tout $\epsilon > 0$ il existe dans R^k un compact K_ϵ tel que $\mathsf{P}((f_{\theta_1}(X), \ldots, f_{\theta_k}(X)) \in K_\epsilon) \geqslant 1 - \epsilon$. Puisque $D \times K_\epsilon$ est un compact dans $D \times R^k$ et que

$$\mathsf{P}(\delta_n \in D, (f_{\theta_1}(X), \ldots, f_{\theta_k}(X)) \in K_\epsilon) \geqslant 1 - \epsilon,$$

la suite de distributions $\Pi_n^{(k)}$ est dense (cf. [5]). Donc, d'après le théorème de Prokhorov [5] il existe une distribution $\overline{\Pi}^{(k)}$ et une sous-suite $n^{(k)} = (n_1^{(k)}, n_2^{(k)}, \ldots)$ telles que $\Pi_{n(k)}^{(k)} \Rightarrow \overline{\Pi}^{(k)}$. Mais les distributions $\overline{\Pi}^{(k)}$ sont visiblement compatibles et par suite le théorème de Kolmogorov affirme qu'il existe sur $(D \times R^T, \sigma_D \times \mathfrak{B}^T)$ une distribution $\overline{\Pi}$ dont $\overline{\Pi}^{(k)}$ sont les restrictions à $(D \times R^k, \sigma_D \times \mathfrak{B}^k)$.

Par ailleurs, on peut admettre que $n^{(k+1)} \subset n^{(k)}$. En posant $n^* = (n_1^{(1)}, n_2^{(2)}, n_3^{(3)}, \ldots)$, on obtient une suite pour laquelle $\Pi_{n^*}^{(k)} \Rightarrow \overline{\Pi}^{(k)}$ pour tous les k.

Prouvons maintenant la relation (12). Supposons que $C \in \mathfrak{B}^T$ est un ensemble cylindrique de frontière $\overline{\Pi}$-négligeable. Désignons par $C^{(k)} = C \cap R^k \in \mathfrak{B}^k$ l'ensemble de R^k formé par les k premières coordonnées des points de C et posons $\overline{C}^{(k)} = C^{(k)} \times R^{T-k} \in \mathfrak{B}^T$. Alors $\lambda(\overline{C}^{(k)}) = \Pi_{n^*}^{(k)}(D \times C^{(k)}) \to \overline{\Pi}^{(k)}(D \times C^{(k)})$. Comme $\overline{C}^{(k+1)} \subset \overline{C}^{(k)}$, $C = \bigcap_{k=1}^{\infty} \overline{C}^{(k)}$, il vient

$$\lambda(C) = \lim_{k \to \infty} \lambda(\overline{C}^{(k)}) = \lim_{k \to \infty} \overline{\Pi}^{(k)}(D \times C^{(k)}) = \lim_{k \to \infty} \overline{\Pi}(D \times \overline{C}^{(k)}) = \overline{\Pi}(D \times C). \quad ◄$$

Table I. Distribution normale réduite $\Phi_{0,1}$

Valeurs de $\overline{\Phi}(x) = \Phi_{0,1}(]x, \infty[) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2/2} dt.$

x	0	1	2	3	4
0,0	0,5000	0,4960	0,4920	0,4880	0,4840
0,1	,4602	,4562	,4522	,4483	,4443
0,2	,4207	,4168	,4129	,4090	,4052
0,3	,3821	,3783	,3745	,3707	,3669
0,4	,3446	,3409	,3372	,3336	,3300
0,5	,3085	,3050	,3015	,2981	,2946
0,6	,2743	,2709	,2676	,2643	,2611
0,7	,2420	,2389	,2358	,2327	,2297
0,8	,2119	,2090	,2061	,2033	,2005
0,9	,1841	,1814	,1788	,1762	,1736
1,0	,1587	,1562	,1539	,1515	,1492
1,1	,1357	,1335	,1314	,1292	,1271
1,2	,1151	,1131	,1112	,1093	,1075
1,3	,0968	,0951	,0934	,0918	,0901
1,4	,0808	,0793	,0778	,0764	,0749
1,5	,0668	,0655	,0643	,0630	,0618
1,6	,0548	,0537	,0526	,0516	,0505
1,7	,0446	,0436	,0427	,0418	,0409
1,8	,0359	,0351	,0344	,0336	,0329
1,9	,0288	,0281	,0274	,0268	,0262
2,0	,0228	,0222	,0217	,0212	,0207
2,1	,0179	,0174	,0170	,0166	,0162
2,2	,0139	,0136	,0132	,0129	,0125
2,3	,0107	,0104	,0102	,0099	,0096
2,4	,0082	,0080	,0078	,0075	,0073
2,5	,0062	,0060	,0059	,0057	,0055
2,6	,0047	,0045	,0044	,0043	,0041
2,7	,0035	,0034	,0033	,0032	,0031
2,8	,0026	,0025	,0024	,0023	,0023
2,9	,0019	,0018	,0018	,0017	,0016
x =	3,0	3,1	3,2	3,3	3,4
$\overline{\Phi}(x)$ =	0,0013	0,0010	0,0007	0,0005	0,0003

Table 1 (*suite*)

x	5	6	7	8	9
0,0	0,4810	0,4761	0,4721	0,4681	0,4641
0,1	,4404	,4364	,4325	,4286	,4247
0,2	,4013	,3974	,3936	,3897	,3859
0,3	,3632	,3594	,3557	,3520	,3483
0,4	,3264	,3228	,3192	,3156	,3121
0,5	,2912	,2877	,2843	,2810	,2776
0,6	,2578	,2546	,2514	,2483	,2451
0,7	,2266	,2236	,2206	,2177	,2148
0,8	,1977	,1949	,1922	,1894	,1867
0,9	,1711	,1685	,1660	,1635	,1611
1,0	,1469	,1446	,1423	,1401	,1379
1,1	,1251	,1230	,1210	,1190	,1170
1,2	,1056	,1038	,1020	,1003	,0985
1,3	,0885	,0869	,0853	,0838	,0823
1,4	,0735	,0721	,0708	,0694	,0681
1,5	,0606	,0594	,0582	,0571	,0559
1,6	,0495	,0485	,0475	,0465	,0455
1,7	,0401	,0392	,0384	,0375	,0367
1,8	,0322	,0314	,0307	,0301	,0294
1,9	,0256	,0250	,0244	,0239	,0233
2,0	,0202	,0197	,0192	,0188	,0183
2,1	,0158	,0154	,0150	,0146	,0143
2,2	,0122	,0119	,0116	,0113	,0110
2,3	,0094	,0091	,0089	,0087	,0084
2,4	,0071	,0069	,0068	,0066	,0064
2,5	,0054	,0052	,0051	,0049	,0048
2,6	,0040	,0039	,0038	,0037	,0036
2,7	,0030	,0029	,0028	,0027	,0026
2,8	,0022	,0021	,0021	,0020	,0019
2,9	,0016	,0015	,0015	,0014	,0014
x =	3,5	3,6	3,7	3,8	3,9
$\overline{\Phi}(x)$ =	0,0002	0,0002	0,0001	0,0001	0,0000

Table II. Quantiles de la distribution normale

Valeurs de λ_ε telles que

$$\overline{\Phi}(\lambda_\varepsilon) = \Phi_{0,1}(]\lambda_\varepsilon, \infty[) = \varepsilon$$

100ε	λ_ε	100ε	λ_ε	100ε	λ_ε
50	0,0000	20	0,8416	0,5	2,5758
45	0,1257	15	1,0364	0,1	3,0902
40	0,2533	10	1,2816	0,05	3,2905
35	0,3853	5	1,6449	0,01	3,7190
30	0,5244	2,5	1,9600	0,005	3,8906
25	0,6745	1	2,3263		

Table III. Distribution H_k du χ^2

Valeurs de (cf. § 2.2)

$$H_k(x) = \mathbf{H}_k(]x, \infty[) = \frac{1}{2^{k/2}\Gamma(k/2)} \int_x^\infty t^{k/2-1} e^{-t/2} dt$$

pour $1 \leqslant k \leqslant 20$. Pour les grands k on peut se servir de l'approximation (cf. § 2.2, table I)

$$H_k(x) \approx \overline{\Phi}(\sqrt{2x} - \sqrt{2k-1}) = \tilde{H}_k(x). \quad (1)$$

La dernière colonne est composée des valeurs de $\tilde{H}_k(x)$ pour $k = 20$. Une comparaison avec la colonne précédente permet d'apprécier le degré de précision de l'approximation (1). L'erreur diminue lorsque k augmente.

x \ k	1	2	3	4	5
0,1	0,7518	0,9512	0,9918	0,9988	0,9998
0,2	,6547	,9048	,9776	,9953	,9991
0,4	,5271	,8187	,9402	,9825	,9953
0,6	,4386	,7408	,8964	,9631	,9880
0,8	,3711	,6703	,8495	,9385	,9770
1,0	,3173	,6065	,8013	,9098	,9626
1,5	,2207	,4724	,6823	,8266	,9131
2	,1573	,3679	,5725	,7358	,8492
3	,0833	,2231	,3916	,5578	,7000
4	,0455	,1353	,2615	,4060	,5494
5	,0254	,0821	,1718	,2873	,4159
6	,0143	,0498	,1116	,1992	,3062
7	,0082	,0302	,0719	,1359	,2206
8	,0047	,0183	,0460	,0916	,1562
9	,0027	,0111	,0293	,0611	,1091
10	,0016	,0067	,0186	,0404	,0752
11	,0009	,0041	,0117	,0266	,0514
12	,0005	,0025	,0074	,0174	,0348
13	,0003	,0015	,0046	,0113	,0234
14	,0002	,0009	,0029	,0073	,0156
15	,0001	,0006	,0018	,0047	,0104
16	,0001	,0003	,0011	,0030	,0068
17		,0002	,0007	,0019	,0045
18		,0001	,0004	,0012	,0030
19		,0001	,0003	,0008	,0019
20		,0001	,0002	,0001	,0013
21			,0001	,0003	,0008
22			,0001	,0002	,0005
23				,0001	,0003
24				,0001	,0002
25				,0001	,0001

Table III (*suite*)

x \ k	6	7	8	9	10
0,5	0,9978	0,9995	0,9999	1,0000	1,0000
1,0	,9856	,9948	,9983	,9994	0,9998
1,5	,9595	,9823	,9927	,9972	,9989
2,0	,9197	,9598	,9810	,9915	,9963
2,5	,8685	,9271	,9617	,9809	,9909
3	,8089	,8850	,9344	,9643	,9814
4	,6767	,7798	,8571	,9114	,9474
5	,5438	,6600	,7576	,8343	,8912
6	,4232	,5398	,6472	,7399	,8153
7	,3204	,4284	,5366	,6371	,7254
8	,2381	,3326	,4335	,5342	,6288
9	,1736	,2527	,3423	,4373	,5321
10	,1246	,1886	,2650	,3505	,4405
11	,0884	,1386	,2017	,2757	,3575
12	,0620	,1006	,1512	,2133	,2851
13	,0430	,0721	,1119	,1626	,2237
14	,0296	,0512	,0818	,1223	,1730
15	,0203	,0360	,0592	,0909	,1321
16	,0138	,0251	,0424	,0669	,0996
17	,0093	,0174	,0301	,0487	,0744
18	,0062	,0120	,0212	,0352	,0550
19	,0042	,0082	,0149	,0252	,0403
20	,0028	,0056	,0103	,0179	,0193
21	,0018	,0038	,0072	,0127	,0211
22	,0012	,0025	,0049	,0084	,0151
23	,0008	,0017	,0034	,0062	,0108
24	,0005	,0011	,0023	,0043	,0076
25	,0003	,0008	,0016	,0030	,0054
26	,0002	,0005	,0011	,0020	,0037
27	,0002	,0003	,0007	,0014	,0026
28	,0001	,0002	,0004	,0010	,0018
29	,0001	,0002	,0003	,0007	,0013
30		,0001	,0002	,0004	,0009

Table III (*suite*)

x \ k	11	12	13	14	15
2	0,9985	0,9994	0,9998	0,9999	1,0000
3	,9907	,9955	,9979	,9991	0,9996
4	,9699	,9834	,9912	,9955	,9977
5	,9312	,9580	,9752	,9858	,9921
6	,8734	,9161	,9462	,9665	,9798
7	,7991	,8576	,9022	,9347	,9577
8	,7133	,7852	,8436	,8893	,9238
9	,6219	,7029	,7729	,8311	,8775
10	,5304	,6160	,6939	,7622	,8197
12	,3636	,4457	,5276	,6063	,6790
14	,2330	,3007	,3738	,4497	,5255
16	,1411	,1912	,2491	,3134	,3821
18	,0816	,1157	,1575	,2068	,2627
20	,0453	,0671	,0952	,1301	,1719
21	,0334	,0504	,0729	,1016	,1368
22	,0244	,0375	,0554	,0786	,1078
23	,0177	,0277	,0417	,0603	,0841
24	,0127	,0203	,0311	,0458	,0651
25	,0091	,0148	,0231	,0346	,0499
26	,0065	,0107	,0170	,0259	,0380
27	,0046	,0077	,0124	,0193	,0287
28	,0032	,0055	,0091	,0142	,0216
29	,0023	,0039	,0066	,0105	,0161
30	,0016	,0028	,0047	,0076	,0119
31	,0011	,0020	,0034	,0055	,0088
32	,0008	,0014	,0024	,0040	,0064
33	,0005	,0010	,0017	,0029	,0047
34	,0004	,0007	,0012	,0021	,0034
35	,0003	,0005	,0009	,0015	,0025
36	,0002	,0003	,0006	,0010	,0018
37	,0001	,0002	,0004	,0007	,0013
38	,0001	,0002	,0003	,0005	,0009
39	,0001	,0001	,0002	,0004	,0006
40		,0001	,0001	,0003	,0005

Table III (*suite*)

k / x	16	17	18	19	20	$H_{20}(x)$
4	0,9989	0,9995	0,9998	0,9999	1,0000	0,9997
5	,9958	,9978	,9989	,9994	0,9997	,9990
6	,9881	,9932	,9962	,9979	,9989	,9973
7	,9733	,9836	,9901	,9942	,9967	,9938
8	,9489	,9666	,9786	,9867	,9919	,9876
9	,9134	,9403	,9597	,9735	,9829	,9774
10	,8666	,9036	,9319	,9530	,9682	,9619
12	,7440	,8001	,8472	,8856	,9161	,9109
14	,5987	,6671	,7291	,7837	,8305	,8298
16	,4530	,5238	,5926	,6573	,7166	,7218
18	,3239	,3888	,4557	,5224	,5874	,5968
20	,2202	,2742	,3328	,3946	,4579	,4683
22	,1432	,1847	,2320	,2843	,3405	,3489
24	,0895	,1194	,1550	,1962	,2424	,2472
26	,0540	,0745	,0998	,1302	,1658	,1670
28	,0316	,0449	,0621	,0834	,1094	,1078
30	,0180	,0264	,0375	,0518	,0699	,0667
31	,0135	,0200	,0288	,0404	,0552	,0517
32	,0100	,0151	,0220	,0313	,0433	,0396
33	,0074	,0113	,0167	,0240	,0337	,0301
34	,0054	,0084	,0126	,0184	,0261	,0227
35	,0040	,0062	,0095	,0140	,0201	,0169
36	,0029	,0046	,0071	,0106	,0154	,0125
37	,0021	,0034	,0052	,0080	,0117	,0092
38	,0015	,0025	,0039	,0059	,0089	,0067
39	,0011	,0018	,0029	,0044	,0067	,0048
40	,0008	,0013	,0021	,0033	,0050	,0035
41	,0006	,0009	,0015	,0024	,0037	,0025
42	,0004	,0007	,0011	,0018	,0028	,0017
43	,0003	,0005	,0008	,0013	,0020	,0012
44	,0002	,0003	,0006	,0009	,0015	,0010
45	,0001	,0002	,0004	,0007	,0011	,0006

Table IV. Distribution $\mathbf{T}_k$ de Student

Valeurs de

$$T_k(x) = \mathbf{T}_k(]x, \infty[) = \frac{\Gamma((k+1)/2)}{\sqrt{k\pi}\,\Gamma(k/2)} \int_x^{\infty} (1 + t^2/k)^{-(k+1)/2} dt$$

pour $1 \leqslant k \leqslant 20$. Pour les k plus grands on peut se servir de l'approximation (cf. table I, § 2.2)

$$T_k(x) \approx \overline{\Phi}(x) = \Phi_{0,1}(]x, \infty[) \qquad (2)$$

La précision de l'approximation (2) pour $k = 20$ peut être appréciée en comparant la dernière colonne de cette table à la table I.

x \ k	1	2	3	4	5
0,0	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	,3524	,3333	,3257	,3217	,3191
1,0	,2500	,2113	,1955	,1869	,1816
1,2	,2211	,1765	,1581	,1482	,1419
1,4	,1974	,1482	,1280	,1170	,1102
1,6	,1778	,1253	,1039	,0924	,0852
1,8	,1614	,1068	,0848	,0731	,0659
2,0	,1476	,0917	,0697	,0581	,0510
2,2	,1358	,0794	,0576	,0463	,0395
2,4	,1257	,0692	,0479	,0372	,0308
2,6	,1169	,0679	,0402	,0300	,0241
2,8	,1092	,0537	,0339	,0244	,0190
3,0	,1024	,0477	,0282	,0200	,0150
3,2	,0964	,0427	,0247	,0165	,0120
3,4	,0910	,0383	,0212	,0136	,0096
3,5	,0862	,0346	,0184	,0114	,0078
3,8	,0819	,0314	,0160	,0095	,0063
4,0	,0780	,0286	,0140	,0081	,0052
4,2	,0744	,0261	,0123	,0068	,0045

Table IV (*suite*)

x \ k	1	2	3	4	5
4,4	0,0711	0,0240	0,0109	0,0058	0,0035
4,6	,0681	,0221	,0097	,0050	,0029
4,8	,0654	,0204	,0086	,0043	,0024
5,0	,0628	,0199	,0077	,0037	,0020
5,2	,0605	,0175	,0069	,0033	,0017
5,4	,0583	,0163	,0062	,0028	,0015
5,6	,0562	,0152	,0056	,0025	,0012
5,8	,0543	,0142	,0051	,0022	,0011
6,0	,0526	,0133	,0046	,0019	,0009
6,2	,0509	,0125	,0042	,0017	,0008
6,4	,0493	,0118	,0039	,0015	,0007
6,6	,0479	,0111	,0035	,0014	,0006
6,8	,0465	,0105	,0033	,0012	,0005
7,0	,0452	,0099	,0030	,0011	,0005
7,2	,0439	,0094	,0028	,0010	,0004
7,4	,0428	,0089	,0025	,0009	,0004
7,6	,0416	,0086	,0024	,0008	,0003
7,8	,0406	,0080	,0022	,0007	,0003
8,0	,0396	,0076	,0020	,0007	,0002

Table IV (*suite*)

x \ k	6	7	8	9	10
0,0	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	,3174	,3162	,3153	,3145	,3139
1,0	,1780	,1753	,1733	,1717	,1704
1,2	,1377	,1346	,1322	,1304	,1289
1,4	,1055	,1021	,0995	,0975	,0959
1,6	,0804	,0768	,0741	,0720	,0703
1,8	,0610	,0574	,0548	,0527	,0510
2,0	,0462	,0428	,0403	,0383	,0367
2,2	,0350	,0319	,0295	,0277	,0262
2,4	,0266	,0237	,0216	,0199	,0186
2,6	,0203	,0177	,0158	,0144	,0132
2,8	,0156	,0132	,0116	,0104	,0094
3,0	,0120	,0100	,0085	,0075	,0067
3,2	,0093	,0075	,0063	,0054	,0047
3,4	,0072	,0057	,0047	,0039	,0034
3,6	,0057	,0044	,0035	,0029	,0024
3,8	,0045	,0034	,0026	,0022	,0017
4,0	,0035	,0026	,0020	,0015	,0013
4,2	,0028	,0020	,0015	,0012	,0009
4,4	,0023	,0016	,0011	,0009	,0007
4,6	,0018	,0012	,0009	,0006	,0005
4,8	,0015	.0010	,0007	,0005	,0004
5,0	,0012	,0008	,0005	,0004	,0003

Table IV (*suite*)

x \ k	11	12	13	14	15
0,0	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	,3135	,3131	,3127	,3124	,3112
1,0	,1694	,1685	,1678	,1671	,1666
1,2	,1277	,1266	,1258	,1250	,1244
1,4	,0945	,0934	,0925	,0916	,0909
1,6	,0689	,0678	,0668	,0660	,0652
1,8	,0496	,0485	,0475	,0467	,0460
2,0	,0354	,0343	,0334	,0326	,0320
2,2	,0250	,0241	,0232	,0225	,0219
2,4	,0176	,0168	,0160	,0154	,0149
2,6	,0123	,0116	,0110	,0105	,0100
2,8	,0086	,0080	,0075	,0071	,0067
3,0	,0060	,0055	,0051	,0048	,0045
3,2	,0042	,0038	,0035	,0032	,0030
3,4	,0030	,0026	,0024	,0022	,0020
3,6	,0021	,0018	,0016	,0014	,0013
3,8	,0015	,0013	,0011	,0010	,0009
4,0	,0010	,0009	,0008	,0007	,0006

Table IV (*suite*)

x \ k	16	17	18	19	20
0,0	0,5000	0,5000	0,5000	0,5000	0,5000
0,5	,3119	,3117	,3116	,3114	,3113
1,0	,1661	,1657	,1653	,1649	,1646
1,2	,1238	,1233	,1228	,1224	,1221
1,4	,0903	,0898	,0893	,0888	,0884
1,6	,0646	,0640	,0635	,0630	,0626
1,8	,0454	,0448	,0443	,0439	,0435
2,0	,0314	,0309	,0304	,0300	,0296
2,2	,0214	,0210	,0205	,0202	,0199
2,4	,0145	,0141	,0137	,0134	,0131
2,6	,0097	,0093	,0090	,0082	,0086
2,8	,0064	,0061	,0059	,0057	,0055
3,0	,0042	,0040	,0038	,0037	,0035
3,2	,0028	,0026	,0025	,0024	,0022
3,4	,0018	,0017	,0016	,0015	,0014
3,6	,0012	,0011	,0010	,0009	,0009
3,8	,0008	,0007	,0007	,0006	,0006
4,0	,0005	,0005	,0004	,0004	,0003

NOTICE BIBLIOGRAPHIQUE

Les commentaires bibliographiques qui suivent sont une tentative de situer l'émergence des idées et résultats fondamentaux développés dans cet ouvrage. Ces commentaires n'ont pas l'ambition d'être exhaustifs et renvoient souvent non point aux articles originaux qui sont d'accès difficile, mais aux manuels, monographies ou articles récapitulatifs qui sont plus faciles à trouver. Des indications bibliographiques et historiques plus détaillées sont données par exemple dans [50] et [91].

Certaines notions fondamentales de statistique mathématique sont nées à l'aube du siècle passé et sont rattachées aux noms de Laplace et Gauss. A la fin du siècle dernier, K. Pearson a inauguré par ses travaux une ère d'intense développement de cette science. Le relais a ensuite été assuré par R. Fisher, J. Neyman, A. Kolmogorov et A. Wald. En Union Soviétique, la statistique mathématique doit ses plus grands progrès à A. Kolmogorov et N. Smirnov.

Chapitre 1

§§ 2, 3, 4. Le théorème de Glivenko-Cantelli a été établi en 1933 (pour une distribution continue, la démonstration revient à Glivenko, pour le cas général, à Cantelli).

La démonstration du théorème 1.2.2 est proche de celle de [53] et constitue un cas particulier d'une approche plus générale basée sur la notion de classe finiment approximable. Cette approche est intégralement développée dans l'Annexe I, où est prouvé le théorème 1.4.2. Une approche analogue a été envisagée indépendamment dans [21]. La loi du logarithme itéré (théorème 1.4.3) est établie dans [45].

§ 6. Les théorèmes 1.6.1 et 1.6.2 relatifs à la distribution de $nF_n^*(t)$ figurent dans l'ouvrage de Feller [26], t. 2, § 3 chap. III. Le théorème 1.6.3. de convergence du processus $\sqrt{n}(F_n^*(t) - F(t))$ vers un pont brownien qui est prouvé dans l'Annexe II a été établi par Donsker dans [22]. Une démonstration du théorème 1.6.3 légèrement différente de celle de l'Annexe II est accessible dans l'ouvrage de Billingsley [5].

§ 7. L'assertion de l'exemple 1.7.3 concernant la distribution limite de la statistique du $\chi^2(X)$ a été obtenue pour la première fois par K. Pearson (cf. [19]).

§ 8. La proposition du corollaire 1.8.2 fait l'objet du théorème de Kolmogorov, et celle du corollaire 1.8.3, du théorème de Smirnov. Ce dernier comprend également la forme explicite de la distribution de $\int_0^1 [w^\circ(t)]^2 dt$, forme qui ne sera pas citée ici en raison de sa complexité (cf. [76]).

§ 10. Les estimations de la densité envisagées dans ce paragraphe ont été introduites par Parzen [64] et Rosenblatt [69]. Les résultats acquis dans cette direction et la bibliographie respective sont accessibles dans le travail récapitulatif de Rosenblatt [70] et dans le § 25 de l'ouvrage de Tchentsov [78].

Chapitre 2

§ 2. D'autres familles paramétriques sont décrites dans l'ouvrage de Wilks [89]. B. Gnédenko a réalisé une étude assez complète des distributions des termes d'un échantillon ordonné. Les résultats et la bibliographie relative à ce sujet sont accessibles dans l'ouvrage de David [20].

§ 4. La méthode des moments est historiquement la première méthode régulière de construction des estimateurs. Elle a été proposée par K. Pearson en 1894.

§ 5. La méthode du minimum du χ^2 a été établie par R. Fisher en 1922.

§ 6. La méthode du maximum de vraisemblance a été utilisée déjà par Gauss dans des cas particuliers. Comme méthode générale de construction d'estimateurs elle a été suggérée par R. Fisher dans une note en 1912. Plus tard, en 1925, Fisher a étudié dans un travail classique [29] les propriétés asymptotiques des estimateurs du maximum de vraisemblance.

§§ 7, 8. Les méthodes de comparaison des estimateurs proposés sont classiques. La démonstration du lemme 2.7.3 a été empruntée à [19]. La notion d'estimateur efficace a été introduite par Fisher en 1922 dans [28].

§§ 9, 10. La notion fondamentale d'espérance mathématique conditionnelle a été proposée par A. Kolmogorov en 1933 dans un travail classique [47]. Les propriétés des distributions conditionnelles sont étudiées en détail dans [17], [24], [34].

§ 11. Le point de vue bayésien était largement utilisé, encore au siècle passé par Laplace. Cette approche a été critiquée par Fisher et dans les années 20 à 30 les recherches ont porté essentiellement sur les estimateurs efficaces et asymptotiquement efficaces. Cette approche fut ensuite remise à l'honneur dès que l'on eut pris conscience de son rôle fondamental.

La notion d'estimateur minimax a été introduite en statistique mathématique en même temps que le point de vue de la théorie des jeux qui a été développé dans les travaux de Borel (1921) et von Neumann (1928) ; les théorèmes 2.11.1, 2.11.2 et 2.11.3 on été prouvés par Hodges et Lehmann [38].

§ 12. La notion fondamentale de statistique exhaustive a été introduite par R. Fisher [28] en 1922. Fisher [28] et plus tard J. Neyman [58] ont proposé un critère simple permettant de déterminer l'existence et la forme d'une statistique exhaustive. Ce critère s'appelle théorème de factorisation de Neyman-Fisher (cf. théorème 2.12.1). Ce théorème n'a été prouvé rigoureusement par les outils de la théorie des ensembles qu'en 1949 par Halmos et Savage [37].

§ 13. La notion de tribu exhaustive est plus large que celle de statistique exhaustive. Les conditions nécessaires et suffisantes de leur coïncidence sont exhibées dans [91]. La détermination des partitions exhaustives et le théorème 2.13.1 sont liés au travail de Lehmann et Scheffe [51] qui est consacré à l'etablissement des conditions d'existence et à la construction des statistiques exhaustives minimales. Cet article est brièvement exposé dans [91]. La démonstration du théorème 2.13.2 appartient à I. Borissov.

§ 14. Le théorème 2.14.1 a été acquis indépendamment par Blackwell [6] en 1947, Rao [67] en 1945, [68] en 1949 et Kolmogorov [46] en 1950. Le théorème 2.14.3 est l'œuvre de Rao [68] (1949) et Blackwell [6] (1947).

§ 15. La famille exponentielle est mentionnée déjà dans les travaux de Fisher [28]. L'importance théorique de cette famille a été appréhendée dans les années 30 par Pitman, Koopman, Darmois. La famille exponentielle porte parfois les noms de ces derniers. Le théorème 2.15.2 a été prouvé par Lehmann [50].

§§ 16, 17. L'inégalité de Rao-Cramer est parfois appelée inégalité d'information. Elle appartient en fait à Fisher [29] bien que dans la forme exhibée elle ait été obtenue indépendamment par Frechet [31] en 1943, Rao [66] en 1945 et Cramer [18] en 1946.

Les conditions de régularité nécessaires à la réalisation de cette inégalité pèchent parfois par leur rigueur dans de nombreux ouvrages de statistique mathématique. Nous avons à l'esprit les conditions assurant la légitimité de la dérivation par rapport au paramètre sous le signe

d'intégration. La démonstration de cette légitimité contient souvent des lacunes (cf. par exemple [91]) ou fait tout simplement défaut (par exemple dans [82]). Dans de nombreux cas, elle est posée comme condition ([82]) ce qui n'est pas commode dans les problèmes d'application.

Les conditions de régularité adoptées dans cet ouvrage sont assez simples même si elles ne sont visiblement pas les plus générales (comparer avec [43]). La possibilité d'une dérivation, dans ces conditions, sous le signe d'intégration est établie dans l'Annexe VI qui s'appuie sur les résultats de A. Sakhanenko.

Diverses généralisations de l'inégalité de Rao-Cramer sont traitées dans [78] et [91]. La notion de quantité d'information (de Fisher) a été introduite dans [29]. Les démonstrations des théorèmes 2.16.1A et 2.17.1 s'inspirent des ouvrages [42] et [91].

§§ 18, 19. L'utilisation de l'invariance est une idée de Hotelling et de Pitman. S. Stein a apporté une importante contribution à l'élaboration de la théorie. Le contenu essentiel du théorème 2.18.1 est dû à Pitman. Pour le prouver, nous nous sommes appuyés sur [42], [91]. La minimaximalité de l'estimateur de Pitman a été établie par Girshik et Savage.

§ 20. Les résultats de ce paragraphe ont été acquis par l'auteur en collaboration avec A. Sakhanenko [13]. Certaines inégalités peuvent être déduites sous des conditions plus restrictives à partir des travaux [32], [77].

§ 21. La distance de Kullback-Leibler dans le cas paramétrique est appelée parfois fonction d'information de Kullback-Leibler. I. Sanov a abouti indépendamment à cette distance en décrivant les probabilités des grands écarts d'une distribution empirique. L'idée d'utiliser largement la distance de Hellinger pour étudier les propriétés du rapport de vraisemblance a été empruntée à Ibraguimov et Khazminski [42]. Cet ouvrage a encore inspiré les démonstrations des théorèmes fondamentaux du § 23. La démonstration du théorème 2.21.3 a été considérablement simplifiée par A. Sakhanenko.

§ 22. Le théorème 2.22.1 a été démontré par Chapman et Robbins [16] en 1951 et par Kiefer [44] en 1952.

§§ 23, 24 et 25. On développe des cours profondément améliorés après la parution de l'ouvrage d'Ibraguimov et Khazminski [42]. Les principaux perfectionnements sont liés à l'utilisation systématique de la distance de Hellinger pour estimer $\mathbf{E}_\theta Z^{1/2}(u)$. L'idée de se servir de $\int \mathbf{E}_\theta |(Z^{3/4}(u))'| du$ pour estimer $\sup_u Z(u)$ (cf. théorème 2.23.1 et 2.23.2) a été avancée par A. Sakhanenko. La normalité asymptotique et l'efficacité asymptotique des estimateurs du maximum de vraisemblance a été établie par Fisher [29]. Des conditions de normalité asymptotique assez générales des estimateurs du maximum de vraisemblance ont été acquises dans [42].

La normalité asymptotique de la densité *a posteriori* (ou du rapport de vraisemblance) a été découverte par S. Berstein en 1927. Le thérème 2.25.4 appartient à Bahadur [1]. On établit sans peine que l'estimateur du maximum de vraisemblance est asymptotiquement bayésien et asymptotiquement minimax grâce aux résultats du § 2.20. On a prouvé antérieurement que l'estimateur du maximum de vraisemblance est asymptotiquement bayésien sous des conditions plus restrictives sur la densité de la distribution *a priori*.

La démonstration des théorèmes 2.24.1 et 2.24.2 utilise quelques perfectionnements proposés par A. Sakhanenko.

§ 26. On développe une variante de la méthode numérique de Newton-Raphson de recherche de l'extremum d'une fonction. Pour un exposé plus détaillé voir [91]. L'exemple 3 a été emprunté à l'ouvrage de Rao [68].

§ 27. L'étude de la convergence de l'estimateur du maximum de vraisemblance a été entamée dans les années 30 et 40 dans les travaux de Doob [23], Wald [84], Wolfowitz [90], Cramer [19]. Les principales conditions de convergence de [84] impliquent (outre les conditions (A_μ), (A_c) et (A_0)) que $f_t(x)$ soit de classe D_0 et que

$$\int \ln f_t^\circ(x) f_\theta(x) \mu(dx)$$

soit intégrable. Dans [42] on établit des conditions de convergence basées sur la convergence

$$\int \sup_{|u| \le \Delta} (\sqrt{f_{t+u}(x)} - \sqrt{f_t(x)})^2 \mu(dx) \to 0 \text{ lorsque } \Delta \to 0.$$

Les résultats des théorèmes 27.1 et 27.2 et leurs corollaires sont plus généraux. Les méthodes de démonstration sont proches de [84]. La suffisance des conditions ($A8$) et (2.27.2) a été remarquée par A. Sakhanenko.

§§ 28, 29. Voir commentaires des §§ 23 à 27. L'exemple 2.28.1 a été emprunté à l'ouvrage de van der Waerden [82]. L'exposé a bénéficié des nombreux perfectionnements proposés par A. Sakhanenko (en particulier le théorème 2.29.5). Ces changements ont permis de simplifier le contenu des §§ 13, 14 et 15 du chapitre 3.

§ 30. Pour de plus amples détails sur l'estimation séquentielle voir par exemple [91].

§§ 31, 32. Les premiers intervalles de confiance font leur apparition dans les travaux de Laplace. Dès 1812 il a montré qu'il était possible d'inverser par rapport à p la proposition concernant le degré de l'écart entre la fréquence observée et la probabilité binomiale p afin de trouver un intervalle pour les valeurs possibles de p. Une interprétation correcte des intervalles de confiance (ne supposant pas la stochasticité du paramètre) a été donnée en 1927 par Wilson.

Une méthode générale de détermination des intervalles de confiance exacts pour un paramètre réel a été proposée par Fisher en 1930 dans [30]. En 1937-1938 Neyman a développé la théorie générale des intervalles de confiance et établi leurs liens avec la théorie de test des hypothèses. Un exposé moderne assez complet de cette question est accessible dans l'ouvrage de Lehmann [50]. Nous nous sommes inspirés de cet exposé dans le § 3.7.

Le théorème 2.32.1 et le lemme 2.32.2 sont l'œuvre de Fisher.

Chapitre 3

Les premières applications éparses des tests statistiques remontent à Laplace (fin du XVIII-ième siècle). L'usage systématique des tests pour éprouver des hypothèses commence avec les travaux de K. Pearson qui a proposé le test du χ^2 en 1900. Les notions fondamentales de risque de première et de deuxième espèce ont été introduites par Neyman et Pearson [60] en 1928. Ces mêmes auteurs ont mis les premiers en évidence le rôle des alternatives pour un choix rationnel du test. La théorie du test uniformément le plus puissant est développée dans le travail de Neyman et Pearson [61].

L'ouvrage de Lehmann [50] expose systématiquement la théorie de test d'hypothèses.

§§ 1, 2 et 3. Le théorème fondamental de Neyman-Pearson est prouvé dans [61]. Les théorèmes 3.1.1 et 3.1.2 figurent dans Blackwell et Girshik [7]. Le théorème 3.2.1 est accessible dans Lehmann [50]. Le théorème 3.3.1 sur les grands écarts est l'œuvre de Cramer (cf. [11]). L'estimation de la qualité d'un test qui est liée aux probabilités des grands écarts sert de base à la notion d'efficacité au sens de Bahadur. Le bilan des recherches effectuées dans cette direction se trouve dans [3].

Le rôle d'une statistique efficace a été signalé dès 1925 par Fisher [29]. Le point de vue lié à l'étude des hypothèses voisines a été intensément développé dans la suite par Le Cam, Roussas, Tchibissov (cf. également les commentaires des §§ 3.14 et 3.15).

§ 4. La conception générale des tests statistiques est passée dans l'usage [cf. [19], [50]). La notion de test uniformément le plus puissant a été introduite par Neyman et Pearson dans [61]. L'approche bayésienne a été utilisée encore au XIX-ième siècle par Laplace.

§§ 5 à 8. Les principaux résultats de ces paragraphes ont été empruntés à Lehmann [50]. L'exposé est fait dans le même esprit que [50] mais sur le point de vue bayésien et non plus sur le lemme généralisé de Neyman-Pearson (lemme 3.5.2, cf. aussi [50]). Ceci simplifie l'exposé et le rend plus cohérent.

Pour les régions de confiance, voir les commentaires des §§ 2.31 et 3.32.

Sur la possibilité de généralisation des résultats fondamentaux à des processus aléatoires voir Grenander [33].

§ 9. Le théorème 3.9.1 a été prouvé par Hodges et Lehmann [38].

§ 10. Le rôle fondamental du rapport de vraisemblance en statistique mathématique est mis en évidence dans les travaux de Neyman et Pearson [60], [61]. Le test du rapport de vraisemblance a fait l'objet de nombreux travaux. Des tentatives pour établir des propriétés d'optimalité asymptotique de ce test ont été effectuées dans les ouvrages [2], [39], [63], [84], [89].

§ 11. La principale contribution à la théorie de l'analyse séquentielle est l'œuvre de Wald [85]. L'exposé le plus condensé des principaux résultats dont nous sommes inspirés est accessible dans [50].

§ 12. Au sujet du test de Kolmogorov et du ω^2, voir § 1.8 et les commentaires respectifs. Sur certaines modifications du test de Kolmogorov conduisant à la plus grande puissance possible, voir [15]. Le test de Moran a été introduit dans [56]. Sa puissance pour des alternatives voisines est étudiée dans [79], [87].

§ 13. Dans [10] on établit que le test du maximum de vraisemblance est asymptotiquement bayésien. Des résultats concernant la distribution limite du rapport de vraisemblance pour l'hypothèse de base ont été acquis par Wilks [88] et Wald [83] (voir également Wilks [89]). L'idée de remplacer une hypothèse multiple par une hypothèse moyennisée a été utilisée par Wald. La forme asymptotique des tests bayésiens figure dans [52]. Cf. également les commentaires des §§ 28 et 29 du chapitre 2.

§§ 14, 15. Les idées fondamentales liées à la recherche des tests asymptotiquement optimaux d'hypothèses voisines sont exposées dans les travaux de Wald [83], Le Cam, Roussas (cf. [71]), Tchibissov [80]. Sur la possibilité de généraliser les résultats fondamentaux au cas d'un paramètre infini (de processus aléatoires), cf. [14]. La forme des exposés des §§ 14 et 15 est peu liée à celle des ouvrages cités. La réduction du problème initial A à un problème B pour le paramètre d'une distribution normale lorsque l'on recherche les tests optimaux pour les principaux types de problèmes envisagés dans le § 14, est accessible dans Wald [83]. Le théorème 3.15.4 relatif à la distribution de la statistique $2 \ln R_1(X)$ pour l'hypothèse H_1 figure dans [89]. Voir également les commentaires des §§ 28, 29 du chapitre 2.

§§ 16, 17. Le test du χ^2 a été proposé par K. Pearson en 1900. Ce test fait l'objet de nombreux travaux (cf. par exemple la monographie spéciale de Lancaster [49]). Les diverses propriétés d'optimalité sont discutées dans [39], [63], [83], [89], etc. Au sujet du comportement de la puissance du test du χ^2 lorsque le nombre de groupes augmente, cf. par exemple [12], [81]. Les exemples 3.16.1 et 3.17.2 ont été empruntés à Cramer [19], l'exemple 3.17.1, à Rao [68].

§ 18. Il est difficile de situer l'origine des recherches entreprises sur la stabilité des décisions statistiques. Des recherches plus tardives sont basées sur les travaux de Tuckey, Hodges et Lehmann. L'ouvrage de Huber [41] donne un aperçu très complet des résultats acquis dans cette direction.

Pour dresser les tables I—IV on s'est inspiré de l'ouvrage de Bolchev et Smirnov [8].

Chapitre 4

§ 1. Le test du χ^2 du problème de l'exemple 4.1.1, le test de Student du problème de l'exemple 4.1.3. et le test de Fisher des problèmes des exemples 4.1.4 et 4.1.5. son très souvent utilisés. Pour les autres propriétés d'optimalité de ces tests, voir Lehmann [50]. L'exemple 4.1.1A a été emprunté à [68]. Le problème de Berens-Fisher fait l'objet de nombreux travaux (cf. [50]).

§ 2. La distribution exacte de la statistique $D_{n,n}$ a été trouvée par Gnédenko et Koroliouk (cf. [26]) ; la distribution limite de la statistique D_{n_1, n_2}, par Smirnov. Le théorème 4.2.2 a été établie pour la première fois dans [55] par la méthode des moments. Sur les tests du signe et de Wilcoxon, voir également [35].

§§ 3, 4. Les problèmes de regression et d'analyse de variance sont développés dans les monographies spéciales de Seber [73] et Scheffe [72]. Voir aussi [19], [50], [68].

§ 5. Dans [10] on trouve une remarque sur l'optimalité asymptotique du test (4.5.3).

Chapitre 5

La théorie des jeux a pris naissance après les travaux de Borel en 1921 et de von Neumann en 1928. Le travail qui a introduit la théorie des jeux en statistique mathématique est le travail classique de Neyman-Pearson [62] qui développait de nombreuses idées fondamentales de la théorie des décisions statistiques. Une contribution importante a été apportée au développement de cette théorie par Wald [86]. La théorie mathématique des jeux est assise dans l'ouvrage de von Neumann et Morgenstern [57].

Un exposé accessible des fondements de la théorie des jeux statistiques figure dans les ouvrages de Blackwell et Girshik [7] et Ferguson [27].

§ 2. L'ouvrage de Mc Kinsey [54] est une introduction assez complète à la théorie des jeux.

§§ 3, 4. Pour une description plux complète des fondements de la théorie des jeux statistiques, voir [7] et [27]. Dans ces deux ouvrages les deux théorèmes fondamentaux de la théorie des jeux statistiques ne sont prouvés que dans le cas particulier où les ensembles D et Θ sont discrets. Ceci s'explique par la complexité du cas général (cf. [86]). En annexe nous avons donné la démonstration la plus facile parmi celles qui sont connues de ce théorème. Cette démonstration a été produite par A. Sakhanenko.

§ 5. Voir commentaires du § 12 chap. 2. Pour plus de détails sur l'absence de biais voir [91].

§ 6. On peut trouver des résultats proches des théorèmes de ce paragraphe dans l'ouvrage d'Ibraguimov et Khazminski [42].

§ 7. Voir commentaires du § 13 chap. 3.

§ 8. Voir commentaires des §§ 14, 15 chap. 2.

Annexe VIII

La démonstration des deux théorèmes fondamentaux de théorie des jeux statistiques est accessible dans [86] et sous des hypothèses plus particulières dans [7], [27]. Dans le présent ouvrage on expose l'approcne proposée par A. Sakhanenko. Les points forts de cette démonstration sont les lemmes 2 et 3. Le lemme 2 n'est pas lié au caractère statistique du jeu; il est basé sur les théorèmes de Hahn-Banach et de Riesz et se rapproche des raisonnements développés par exemple dans [25]. La démonstration du lemme 3 repose sur les théorèmes de Kolmogorov [47] et de Prokhorov [5].

BIBLIOGRAPHIE

1. BAHADUR R. R. On Fisher's bound for asymptotic variances. — Ann. Math. Statis., 1964, 35, 4, 1545—1552.

2. BAHADUR R. R. An optimal property of the likelihood ratio statistic, Proc. 5-th Berkeley Sympos. Math. Statist. Prob. — Berkeley Los Angeles, v. 1, 1965, 27—40.

3. BAHADUR R. R. Some limit theorems in statistics. — Philadelphia : S.I.A.M., 1971.

4. BAHADUR R. R., LEHMAN E. L. Two comments on « Sufficiency and statistical decision functions ». — AMS, 1955, 26, 139—141.

5. BILLINGSLEY P. Convergence of probability measures. — John Wiley, New York, 1968.

6. BLACKWELL D. Conditional expectation and unbiased sequential estimation. — Ann. Math. Statist., 1947, 18, 105—110.

7. BLACKWELL D., GIRSHIK M. Theory of Games and Statistical Decisions. — John Wiley, New York, 1954.

8. BOLCHEV L., SMIRNOV N. Tables de statistique mathématique. — M.: Naouka, 1965 (en russe).

9. BOROVKOV A. Stochastic Processes in Queueing Theory. Springer Verlag 1976,

10. BOROVKOV A. Tests asymptotiquement optimaux d'hypothèses multiples. — Teoria veroïatnosteï i eïe primenenia, 1975, 20, 3, 463—487 (en russe).

11. BOROVKOV A. Probability Theory. Birkhäser Verlag, 1976.

12. BOROVKOV A. Sur la puissance du test du χ^2 lorsque le nombre de groupes croît. — Teoria veroïatnosteï i eïe primenenia, 1977, 22, 2, 375—379 (en russe).

13. BOROVKOV A., SAKHANENKO A. Inégalités de type Rao-Cramer pour le risque bayésien. — Teoria veroïatnosteï i eïe primenenia, 1980, 25, 1, 207—209 (en russe).

14. BOROVKOV A., SAKHANENKO A. Sur les tests asymptotiquement optimaux d'hypothèses multiples. — Troudy Ins-ta matematiki SO AN SSSR, 1981, t. 1 (en russe).

15. BOROVKOV A., SYTCHEVA N. Sur certains tests non paramétriques asymptotiquemment optimaux. — Teoria veroïatnosteï i eïe primenenia, 1968, 13, 3, 385—418 (en russe).

16. CHAPMAN D. G., ROBBINS H. E. Minimum variance estimation without regularity assymptions. — Ann. Math. Statist., 1951, 22, 581—586.

17. CHIRIAEV A. Probabilité. — M.: Naouka, 1980.

17.* COX D. R., HINKLEY D. Theoritical statistics. — Chapman Hill, London, 1974.

18. CRAMER H. A contribution to the theory of statistical estimation. — Aktuariestidskrift, 1946, 29, 458—463.

19. CRAMER H. Mathematical Methods of Statistics — Princeton University Press, 1946.

20. DAVID H. A. Ordre statistics. — New York (a.o), cop. 1970.

21. DE HARDT J. Generalizations of the Glivenko-Cantelli theorem. — Ann. Math. Stat., 1971, 42, 2050—2055.

22. DONSKER M. Justifications and extensions of Doob's heuristic approach to the Kolmogorov—Smirnov theorems. — Ann. Math. Statist., 1952, 23, 277—281.

23. DOOB J. L. Probability and statistics. — Trans. Amer. Math. Soc., 1934, 36, 4, 759—775.

24. DOOB J. Stochastic Processes. — John Wiley, New York, 1953.

25. EDWARDS R. Functional Analysis. Theory and Applications. — New York, 1965.

26. FELLER W. An Introduction to Probability Theory and its Applications. — New York, 1966.

27. FERGUSON T. S. Mathematical statistics. A decision theoretic approach. — New York and London: Academic Press, 1967.

28. FISHER R. A. On the mathematical foundations of theoretical statistics. — Phil. Trans. Roy. Soc. A, 1922, 222, 309-368.

29. FISHER R. A. Theory of statistical estimation. — Proc. Camb. Phil. Soc., 1925, 22, 700-725.

30. FISHER R. A. Inverse probability. — Proc. Cambridge Phil. Soc., 1930, 26, 528-535.

31. FRECHET M. Revue internationale de statistique. 1943, 182.

32. GOUSSEV S. Développements asymptotiques liés à certaines estimations statistiques dans le cas régulier. II. — Teoria veroïatnosteï i eïe primenenia. — 1976, 21, 1, 16-33.

33. GRENANDER U. Stochastic Processes and Statistical Inference. — Almqvist and Wiksells Boktryckeri A. B., 1950.

34. GUIKHMAN I., SKOROKHOD A. Introduction à la théorie des processus aléatoires. — M.: Editions Mir, 1980 (Traduction française).

35. HAJEK J., SIDAK Z. Theory of rank tests. — Prague, « Academia », 1967.

36. HALMOS P. R. Measury Theory. — Van Nostrand, Princeton, 1950.

37. HALMOS P. R., SAVAGE L. J. Application of the Radon-Nikodym theorem to the theory of sufficient statistics. — Ann. Math. Statist., 1949, 20, 225-241.

38. HODGES J., LEHMANN E. Some problems in minimax estimation. — Ann. Math. Statist., 1950, 21, 2, 182-197.

39. HOEFFDING W. Asymptotically optimal tests for multinomial distributions. — Ann. Math. Statist., 1965, 36, 2, 369-401.

40. HOTELLING H. The generalization of student's ratio. — Ann. Math. Statist., 1931, 2, 360-378.

41. HUBER P. J. Robust statistics: a review. — Ann. Math. Statist., 1972, 43, 1041-1067.

42. IBRAGUIMOV I., KHAZMINSKI R. Théorie de l'estimation asymptotique. — M.: Naouka, 1979 (en russe).

43. KENDALL M. and STUART A. Inference and Relationship. — (2nd ed), Charles Griffin and Company Limited, London, 1967.

43.* KENDALL M. and STUART A. The advanced theory of statictics. — v. 2, Charles Griffin and Company Limited, London, 1961.

44. KIEFER J. On minimum variance estimators. — Ann. Math. Statist., 1952, 23, 627—629.

45. KIEFER J. On large deviations of the identically distributed functions of vector chance variables and LIL. — Pacif. J. Math., 1961, 11, 2, 649-660.

46. KOLMOGOROV A. Estimateurs sans biais. — Izv. AN SSSR, ser. mat., 14, 1950, 303 (en russe).

47. KOLMOGOROV A. Notions fondamentales de théorie des probabilités. — M.: Naouka, 1974 (en russe).

48. KULLBACK S., LEIBLER R. A. On information and sufficiency. — Ann. Math. Statist., 1951, 22, 79-86.

49. LANCASTER H. O. The chi-squared distribution. — New York: John Wiley and Sons, 1969.

50. LEHMANN E. L. Testing Statistical Hypotheses. — John Wiley, New York, 1959.

50.* LEHMANN E. L. Theory of point estimation. — John Wiley, New York, 1983.

51. LEHMANN E.L., SCHEFFE H. Completeness, similar regions and inbiased estimation. — Pt. I. Sankhya, 1950, **10**, 305-340.

52. LINDLEY D. The use of prior probability distributions in statistical inference and decision. Proc. 4-th Berkeley Sympos. Math. Statist. Prob., Berkeley — Los Angeles, v. 1, 1960, 453-468.

53. LOÈVE M. Probability Theory. — (2nd ed), Van Nostrand, Princeton, 1960.

54. MC KINSEY J. C. Introduction to the Theory of Games — th. ed. New York, a.o. Mc. Graw-Hill, 1952.

55. MANN H. B., WHITNEY D. R. On a test whether one of two random variables is stochastically larger than the other. — Ann. Math. Statist., 1947, **18**, 50.

56. MORAN P. A. P. The random division of an interval. — J. Roy. Stat. Soc., Suppl., 1947, **9**, 92-98.

57. VON NEUMANN J. and MORGENSTERN O. Theory of games and economic behaviour. — Princeton University Press, 1944.

58. NEYMAN J. Sur un theorems concerente le cosidette statistiche sufficienti. — Inst. Ital. Atti. Giorn., 1935, **6**, 320-334.

59. NEYMAN J. First course in probability and statistics. — Holt Rinehart and Winston Jnc, New York, 1951.

60. NEYMAN J., PEARSON E. S. On the use and interpretation of certain test criteria. — Biometrika, 1928, **20A**, 175-240, 263-294.

61. NEYMAN J., PEARSON E. S. On the problem of the most efficient tests of statistical hypotheses. — Phil. Trans. Roy. Soc., Ser. A, 1933, **231**, 289-337.

62. NEYMAN J., PEARSON E. S. The testing of statistical hypotheses in relation to probabilities a priori. — Proc. Camb. Phil. Soc. 1933, **24**, 492-510.

63. OOSTERHOFF J., W. R. VAN ZWET. The likelihood ratio test for the multinomial distribution. Proc. 6-th Berkeley Sympos. Math. Statist. Prob., Berkeley — Los Angeles, v. 1, 1970, 31-50.

64. PARZEN E. On estimation of a probability density function and mode. — Ann. Math. Statist., 1962, **33**, 3, 1065—1076.

65. PITMAN E.J.G. The estimation of the location and scale parameters of a continuous population of any given form. — Biometrika, 1938, **30**, 391-421.

66. RAO C.R. Information and accuracy attainable in estimation of statistical parameters. — Bull. Calcutta Math. Soc., 1945, **37**, 81-91.

67. RAO C.R. Sufficient statistics and minimum variance estimates. — Proc. Cambr. Phil. Soc., 1949, **45**, 213-218.

68. RAO C. R. Linear statistical inference and its applications. — J. Wiley, New York, London, Sidney, 1973, 2nd ed.

69. ROSENBLATT M. Remarks on some non-parametric estimates of a density function. — Ann. Math. Statist., 1956, **27**, 3, 832-837.

70. ROSENBLATT M. Curve estimation. — Ann. Math. Statist., 1971, **42**, 6, 1815-1842.

71. ROUSSAS G. Contiguity of Probability Measures. — Cambridge, 1972.

72. SCHEFFE H. The analysis of variance. — Wiley, Lnd., Chapman & Hall, cop. 1959.

73. SEBER G. A. Linear Regression Analysis. — John Wiley, New York, 1977.

74. SIDOROV V. A., . . . Measurement of the $\varphi \to \pi^+\pi^-$ branching ratio. — Physics Letters, 1981, **99 B**, 1, 62—65.

75. SKOROKHOD A. Processus aléatoires à accroissements indépendants. — M.: Naouka, 1964 (en russe).

76. SMIRNOV N. Sur la distribution du test du ω^2 de Mises. — Dans l'ouvrage: SMIRNOV N. Théorie des probabilités et statistique mathématique. Izbrany troudy. — M.: Naouka, 1970 (en russe).

77. TCHENTSOV N. Sur l'estimation de la moyenne inconnue d'une loi normale multidimensionnelle. — Teoria veroïatnosteï i eïe primenenia, 1967, **12**, 4, 619—633 (en russe).

78. TCHENTSOV N. Décisions statistiques et inférences optimales. — M.: Naouka, 1972 (en russe).

79. TCHIBISSOV D. Sur les tests d'ajustement basés sur des intervalles empiriques. — Teoria veroïatnosteï i eïe primenenia, 1961, **6**, 1, 354-358 (en russe).

80. TCHIBISSOV D. (CHIBISOV D.). Transition to the limiting process for deriving asymptotically optimal tests. — Sankhya, 1969, **A 31**, 3, 241-258.

81. TCHIBISSOV D., GVANTSELADZE L. Sur les tests d'ajustement basés sur les données groupées. — Dans l'ouvrage: III Symposium soviéto-japonais de théorie des probabilités. — Tachkent: Fan, 1975, 183-185.

82. VAN DER WAERDEN. Statistique mathématique. — Trad. par M-me C. Guinchat. A. Degenne, Paris, Dunod, 1967.

83. WALD A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. — Trans. Amer. Math. Soc., 1943, **54**, 3, 426-482.

84. WALD A. Note on the consistency of the maximum likelihood estimate. — Ann. Math. Statist., 1949, **20**, 595-601.

85. WALD A. Sequentiel Analysis. — N. Y. Wiley, Lnd, Chapman & Hall, 1947.

86. WALD A. Statistical decision functions. — New York, 1950.

87. WEISS L. The asymptotic power of certain tests of fit based on sample spacings. — Ann. Math. Statist., 1957, **28**, 3, 783-786.

88. WILKS S. S. The large sample distribution of the likelihood ratio for testing composite hypotheses. — Ann. Math. Statist., 1938, **9**, 60-62.

89. WILKS S. S. Mathematical Statistics. — John Wiley, New York, 1962.

90. WOLFOWITZ J. On Wald's proof of the consistency of the maximum likelihood estimate. — Ann. Math. Statist., 1949, **20**, 601—602.

91. ZACKS SH. The theory of Statistical Inference. — New York, 1971.

LISTE DES PRINCIPALES NOTATIONS

(A_0)	condition de correspondance biunivoque entre l'ensemble des paramètres et la famille des distributions $\mathscr{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta}$ ($\mathbf{P}_{\theta_1} \neq \mathbf{P}_{\theta_2}$ si $\theta_1 \neq \theta_2$)
(A_c)	condition de compacité de l'ensemble Θ des paramètres
(A_μ)	condition que les distributions de la famille $\mathscr{P} = \{\mathbf{P}_\theta\}$ sont dominées par la mesure μ (la densité $f_\theta = d\mathbf{P}_\theta/d\mu$ existe)
b, $b(\theta)$	biais
$\mathfrak{B}$	tribu des boréliens de la droite R
$\mathfrak{B}_{\mathscr{X}}$	tribu sur l'espace des phases $\mathscr{X}$ (des boréliens si $\mathscr{X} = R^m$)
$\mathbf{B}_p$	distribution polynomiale (y compris la distribution de Bernoulli)
$C(a,b)$	espace des fonctions continues sur $[a, b]$
D	espace des stratégies du premier joueur
$\mathscr{D}$	espace des décisions dans un jeu statistique
$D(a, b)$	espace des fonctions sur $[a, b]$ continues à gauche (au point a à droite) et présentant un nombre fini de sauts
E	matrice unité
$\mathscr{E}$	famille exponentielle de distributions
$\mathbf{E}_\theta$	espérance mathématique par rapport à la distribution $\mathbf{P}_\theta$
$\mathbf{E}(\xi\|\eta)$	espérance mathématique conditionnelle de ξ par rapprort à la variable aléatoire η
$\mathbf{E}(\xi\|\mathfrak{A})$	espérance mathématique conditionnelle de ξ par rapport à la tribu $\mathfrak{A}$
$f_\theta(x)$	densité de la distribution $\mathbf{P}_\theta$ par rapport à la mesure μ
$f_\theta(X) = \prod_{i=1}^{n} f_\theta(x_i)$	fonction de vraisemblance
$F(x)$	fonction de répartition de la distribution $\mathbf{P}$
$F_n^*(x)$	fonction de distribution empirique
$\mathbf{F}_{k_1,k_2}$	distribution de Fisher
G	groupe des transformations de $\mathscr{X}$ dans lui-même, associé à une famille invariante
h_ε	quantile d'ordre $1 - \varepsilon$ de la distribution du χ^2
H_i	hypothèse
$\mathbf{H}_k$	distribution du χ^2
$\mathbf{I}_x$	distribution concentrée au point x
$I(\theta)$	quantité d'information de Fisher
I_A	indicateur de l'ensemble A
K_b	classe des estimateurs de biais $b = b(\theta)$
K_0	classe des estimateurs sans biais
$\tilde{K}_0$	classe des estimateurs asymptotiquement sans biais

$\bar{K}^{\circ}$	classe des estimateurs asymptotiquement centrés
$K_{\Phi,2}$	classe des estimateurs θ^* asymptotiquement normaux pour lesquels $\mathbf{E}_{\theta}n(\theta^* - \theta)^2 \to \sigma^2(\theta)$, où $\sigma^2(\theta)$ est la variance de la distribution normale limite de $\sqrt{n}(\theta^* - \theta)$
K_{ϵ}	classe des tests de dimension ϵ (de niveau $1 - $)
$\bar{K}_{\epsilon}$	classe des tests de niveau asymptotique $1 - \epsilon$
$K_{\epsilon Q'}$	classe des tests de niveau $1 - \epsilon$ pour l'approche partiellement bayésienne
$\bar{K}_{\epsilon Q'}$	classe des tests de niveau asymptotique $1 - \epsilon$ pour l'approche partiellement bayésienne
$K_{\alpha_1, \ldots, \alpha_{r-1}}$	classe des tests à risques α_i de i-ième espèce fixes, $i = 1, \ldots, r - 1$
$\mathbf{K}_{\alpha,\sigma}$	distribution de Cauchy
$l(x, \theta) = \ln f_{\theta}(x)$	
$L(x, \theta) = \ln f_{\theta}(X)$	logarithme de la fonction de vraisemblance
$\mathbf{L}_{\alpha,\sigma^2}$	distribution lognormale
n	taille de l'échantillon
$N_{\mathbf{P}}, N_F$	support de la distribution $\mathbf{P}$ de fonction de répartition F
$\mathbf{P}$	symbole de la distribution utilisé dans les sens indiqués à la page
$\mathbf{P}_n^*$	distribution empirique
$\mathbf{P}_{\theta}$	distribution dépendant du paramètre θ
$P(B\|y)$	distribution conditionnelle
$\mathcal{P}$	famille de distributions
$\mathbf{Q}$	stratégie randomisée de la «nature» (distribution *a priori* de θ)
$\mathbf{Q}_x$	distribution *a posteriori*
$\mathbf{Q}$	distribution la plus défavorable de θ (stratégie minimax de la nature)
$q(t\|X)$	densité de la distribution *a posteriori*
R	droite réelle
R^m	espace euclidien à m dimensions
(R)	conditions de régularité d'une famille paramétrique en vertu desquelles la fonction $\sqrt{f_{\theta}(x)}$ est continûment dérivable par rapport à θ et la quantité d'information de Fisher est strictement positive et continue
(RR)	conditions de régularité d'une famille paramétrique exigeant que les conditions (A_0), (A_c) et (R) soient remplies, que la fonction $l(x, \theta)$ admette des dérivées première et seconde continues, et un majorant $l(x) \geqslant \|l''(x, t)\|$ tel que l'intégrale $\mathbf{E}_{\theta}l(\mathrm{x}_1)$ converge uniformément sur Θ
$S = S(X)$	statistique
S^2	variance empirique
$S_0^2 = \dfrac{n}{n-1} S^2$	
$\mathbf{T}_k$	distribution de Student
$\mathbf{U}_{a,b}$	distribution uniforme sur $[a, b]$
$u^* = \sqrt{n}(\hat{\theta}^* - \theta)$	estimateur normal du maximum de vraisemblance
$\mathbf{V}_{\theta}$	variance par rapport à la distribution $\mathbf{P}_{\theta}$
$w(t)$	processus wienérien (pas partout)
$w^{\circ}(t)$	pont brownien
$w^n(t)$	processus empirique
$w(\delta, \theta)$	fonction de perte du premier joueur
$W(\delta(\cdot), \theta) = \mathbf{E}_{\theta}w(\delta(X), \theta)$	fonction de risque
x_i	élément d'un échantillon

$X = X_n = (x_1, \ldots, x_n)$	échantillon de taille n
$[X_\infty]_n = X_n$	partie d'un échantillon infini, composée des n premiers éléments
$x_{(i)}$	i-ième élément de l'échantillon ordonné (série variationnelle)
$\bar{x}$	moyenne empirique
$\mathscr{X}$	espace des observations (espace des phases de l'échantillon)
$(\mathscr{X}, \mathfrak{B}_\mathscr{X}; \mathbf{P})$	espace probabilisé des échantillons associé à une seule observation
$(\mathscr{X}^n, \mathfrak{B}^n_\mathscr{X}; \mathbf{P})$	espace probabilisé des échantillons associé à un échantillon de taille n
$x = (x_1, \ldots, x_n)$	élément de $\mathscr{X}^n$
$\alpha_i(\pi)$	risque de i-ième espèce du test π
$\beta(\delta)$	fonction de puissance du test δ
$\beta_\pi(\theta)$	fonction de puissance du test π
$\mathbf{B}_{\lambda_1, \lambda_2}$	distribution bêta
$\Gamma_{\alpha,\lambda}$	distribution gamma
δ	stratégie du premier jouer
$\delta = \delta(X)$	décision (test)
ζ_p	quantile d'ordre p
ζ_p^*	quantile empirique d'ordre p
θ	paramètre, stratégie de la nature
θ^*	borne de l'intervalle de confiance pour le paramètre θ
θ^*	estimateur du paramètre θ
$\theta^*_\mathfrak{Q}$	estimateur bayésien du paramètre θ associé à la distribution *a priori* Q
$\bar{\theta}^*$	estimateur minimax du paramètre θ
$\hat{\theta}^*$	estimateur du maximum de vraisemblance du paramètre θ
Θ	ensemble des valeurs possibles du paramètre θ, espace des stratégies de la nature
Θ^*	ensemble de confiance
λ_ε	quantile de la distribution normale
π	stratégie randomisée du premier joueur
$\pi = \pi(X)$	test randomisé, décision randomisée
π_Q	test bayésien associé à la distribution *a priori* Q, stratégie randomisée
$\pi_{\mathrm{Q},\mathfrak{Q}}$	test bayésien pour l'approche partiellement bayésienne
$\bar{\pi}$	test minimax, stratégie minimax
$\hat{\pi}$	test du rapport de vraisemblance
π°	test uniformément le plus puissant
Π_λ	distribution de Poisson
Φ_{α,σ^2}	distribution normale
$\Phi(x)$	fonction de répartition de la loi normale réduite
$\overset{=}{d}$	coïncidence des distributions des échantillons ou des variables aléatoires
$\underset{P}{\to}$	convergence en probabilité
$\underset{p.s.}{\to}$	convergence presque sûre (avec la probabilité 1)
$\Rightarrow$	convergence faible des variables aléatoires ou des distributions
$\in$	relie un échantillon ou une variable aléatoire à une distribution et exprime que cet échantillon ou cette variable sont distribués suivant cette loi
$\Subset$	convergence faible. La relation $\xi_n \Subset \mathbf{P}$ exprime que la distribution de ξ_n converge faiblement vers $\mathbf{P}$ lorsque $n \to \infty$
$\varepsilon_n \uparrow \varepsilon$	ε_n tend par valeurs croissantes vers ε.
$\varepsilon_n \downarrow \varepsilon$	ε_n tend par valeurs décroissantes vers ε.

INDEX

À NOS LECTEURS

Les Editions Mir vous seraient très reconnaissantes de bien vouloir leur communiquer votre opinion sur le contenu de ce livre, sa traduction et sa présentation, ainsi que toute autre suggestion.

Notre adresse:
Editions Mir,
2, Pervy Rijski péréoulok,
Moscou, I-110, GSP, U.R.S.S.

Imprimé en Union Soviétique